RADEMUON: PROVABLE RESOURCE-AWARE QUANTIZED DECENTRALIZED MUON VIA FINITE-STEP NEWTON–SCHULZ ORTHOGONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributed Muon optimizer enables matrix-aware optimization across multiple devices, but its full-precision local training and model exchange impose substantial resource costs, particularly when resource availability changes throughout decentralized training. We propose *RADeMuon*, a resource-aware quantized framework for decentralized Muon that supports precision configurations varying with available resources. *RADeMuon* performs local training with quantized parameters, gradients, and momentum states, applies finite-step Newton–Schulz (NS) orthogonalization to the quantized momentum, and directly communicates the resulting quantized parameter for decentralized aggregation. Theoretically, we trace quantization errors from their accumulation in the momentum state to their transformation by finite-step NS and subsequent propagation through decentralized aggregation. This analysis establishes a convergence guarantee whose dominant stationarity term is $\mathcal{O}((nKT)^{-1/4})$ under vanishing-error conditions. Experiments on image classification and language-model pretraining show that *RADeMuon* reduces local computation and communication costs by up to 58.34% and 58.67%, respectively, relative to full-precision De-EF21-Muon while maintaining comparable model quality.

1 INTRODUCTION

Muon has recently been extended to distributed model training (Liu et al., 2025; Ahn et al., 2025; Gruntkowska et al., 2026; He et al., 2025). Muon maintains matrix-valued momentum and transforms it toward its polar direction before updating the parameters. Practical implementations approximate this transformation using a small number of Newton–Schulz (NS) iterations instead of an exact SVD. In distributed training, repeatedly applying NS and maintaining full-precision states increase local resource consumption, while exchanging full-precision model information introduces substantial communication cost. These overheads hinder the deployment of distributed Muon in resource-constrained environments. Quantization lowers the precision of local training states and reduces communication costs, motivating the study of quantized distributed Muon.

Resource availability in practical decentralized systems differs among devices and changes throughout training. A fixed numerical precision cannot consistently match the changing resource budgets: high precision may be infeasible under tight constraints, whereas uniformly low precision introduces avoidable quantization error when more resources are available. Existing distributed quantization methods reduce the cost of gradient or model communication (Alistarh et al., 2017; Koloskova et al., 2019), while Adriana adjusts local-training and communication precision according to resource availability (Qiao et al., 2026). However, these methods are designed primarily for SGD-type optimization and do not account for Muon's stored matrix momentum and subsequent orthogonalization. Existing low-precision Muon methods, in contrast, mainly consider fixed precision in single-device training (Gupta et al., 2025; Wu et al., 2026; Su et al., 2026). We therefore study a resource-aware quantized decentralized Muon framework that supports precision configurations following the available resources throughout training.

Allowing quantization precision to vary throughout decentralized training creates a Muon-specific theoretical challenge with three coupled layers. **(i) Temporal accumulation of momentum errors.**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison with representative Muon-based and quantized or compressed distributed optimization methods. Rates are rewritten in the notation of this paper.

Method	Setting	Muon	Finite NS	Quant./Comp.	Time-varying Topo.	Stationarity measure*	Rate**
Muon-NS (Kim & Oh, 2026)	Centralized	✓	✓	×	–	$\mathbb{E}\,\|\nabla f(X)\|_*$	$\chi_J\,\mathcal{O}(T^{-1/4})$
FedMuon (Takezawa et al., 2025)	Federated	✓	✓	×	×	$\mathbb{E}\,\|\nabla f(X)\|_{p_J}$	$\mathcal{O}((nKT)^{-1/4})^{\dagger}$
DeMuon (He et al., 2025)	Fully decentralized	✓	×	×	×	$\mathbb{E}\,\|\bar{g}(\mathbf{X})\|_*$	$\mathcal{O}(T^{-1/4})$
Adriana (Qiao et al., 2026)	Fully decentralized	×	–	✓	✓	$\mathbb{E}\,\|\nabla f(\bar{\theta})\|_2^2$	$\mathcal{O}((nKT)^{-1/2}) + C_1^{\ddagger}$
EF21-Muon (Gruntkowska et al., 2026)	Parameter server	✓	×	✓	×	$\mathbb{E}\,\|\nabla f(X)\|_*$	$\mathcal{O}(T^{-1/4})$
RADeMuon (Ours)	Fully decentralized	✓	✓	✓	✓	$\mathbb{E}\,\|\nabla f(\bar{X})\|_*$	$\chi_J\left[\mathcal{O}((nKT)^{-1/4}) + C_2^{\ddagger}\right]$

* $\|\cdot\|_*$ denotes the nuclear norm, $\|\cdot\|_\star$ a general dual norm, and $\|\cdot\|_{p_J}$ the Schatten norm induced by J finite NS steps; $\bar{g}(\mathbf{X}) = n^{-1}\sum_i \nabla f_i(X_i)$.
** n, T, and K denote the number of clients, communication rounds, and local steps per round, respectively. For finite-NS methods, ϵ_J denotes the polar-approximation error after J NS steps, and $\chi_J := (1-\epsilon_J)^{-1}$ denotes the corresponding multiplicative factor.
† Only the dominant term is shown; lower-order terms are omitted.
‡ C denotes the non-vanishing residual induced by stochasticity and data heterogeneity.

Gradient quantization perturbs each momentum update, while quantizing the stored momentum state propagates the accumulated error through subsequent steps and communication rounds. The resulting error therefore depends on the history of precision configurations rather than only on the current quantization. **(ii) Coupling between momentum quantization and finite-step NS.** The finite-step NS map is a nonlinear polynomial applied directly to the quantized momentum. Its output error combines the accumulated momentum error with the approximation residual caused by terminating orthogonalization after only J iterations. This interaction cannot be captured by replacing NS with an exact SVD-based polar factor. **(iii) Error propagation through decentralized aggregation.** Each local update operates on a quantized parameter state, and the resulting quantized parameters are subsequently used in decentralized aggregation. Parameter-quantization errors therefore accumulate along local trajectories, propagate through the communication graph, and interact with consensus disagreement. Existing finite-NS analyses focus on centralized Muon (Kim & Oh, 2026), whereas convergence analyses of fully decentralized Muon typically rely on exact SVD-based polar updates (He et al., 2025). This raises a central question: *Can decentralized Muon support precision configurations that follow changing resource availability while retaining a provable convergence guarantee under practical finite-step NS?*

To answer this question, we propose *RADeMuon*, a resource-aware quantized framework for fully decentralized Muon under heterogeneous and time-varying resource constraints. The framework accepts a feasible precision configuration for each client and communication round, allowing the parameter, gradient, and momentum widths to follow the available local and communication budgets. Clients perform local training with finite-step NS orthogonalization and directly exchange the resulting quantized parameter representation with their neighbors. Theoretically, we establish a finite-horizon convergence bound for *RADeMuon* that characterizes how quantization levels, network connectivity, and finite-step NS accuracy affect decentralized Muon. When the descent margin remains positive and the residual terms vanish, its leading stationarity term is $\mathcal{O}((nKT)^{-1/4})$. Our contributions are summarized as follows:

- To the best of our knowledge, we propose *RADeMuon*, the first resource-aware quantized framework for fully decentralized Muon that supports precision configurations following the available resources throughout training. The quantized parameters produced by local training are communicated directly to neighboring clients.
- We establish a finite-horizon convergence bound for decentralized Muon under heterogeneous and time-varying quantization and reveal how quantization, network connectivity, and finite-step NS affect its convergence. Under a positive descent margin and vanishing residual terms, the leading stationarity term is $\mathcal{O}((nKT)^{-1/4})$.
- We conduct extensive experiments on image classification and language-model pretraining. Compared with full-precision De-EF21-Muon, *RADeMuon* reduces local computation and communication costs by up to 58.34% and 58.67%, respectively, while maintaining comparable model quality.

2 RELATED WORK

Quantization and compression have been widely studied to improve the efficiency of distributed learning. QSGD investigates unbiased gradient quantization and its communication–variance tradeoff (Alistarh et al., 2017), CHOCO studies compressed decentralized optimization (Koloskova et al., 2019), error feedback compensates for information loss caused by gradient compression (Karimireddy et al., 2019), and PowerSGD communicates low-rank gradient factors (Vogels et al., 2019). More recently, Adriana (Qiao et al., 2026) introduces resource-aware quantization into decentralized learning by adapting local-training and communication precision to heterogeneous resources. In parallel, several studies have explored efficient and distributed Muon optimization. Quantized Muon (Gupta et al., 2025) investigates low-bit optimizer states, EF21-Muon (Gruntkowska et al., 2026) employs error-feedback compression, and Dion (Ahn et al., 2025) uses amortized power iteration for sharded orthonormalized updates. DeMuon (He et al., 2025) and SUDA-Muon (Zhang et al., 2026) further extend Muon to decentralized optimization. Existing distributed quantization methods are primarily designed for SGD-type updates, while existing Muon methods do not jointly analyze precision configurations that follow changing resource availability, finite-step NS, and decentralized aggregation.

3 METHOD

3.1 PROBLEM SETUP

We consider a fully decentralized learning system consisting of n clients that collaboratively train a model without a central parameter server. In communication round t, the network topology is represented by an undirected graph $\mathcal{G}^t = (\mathcal{V}, \mathcal{E}^t)$, where $\mathcal{V} = \{1, \dots, n\}$ denotes the set of clients. The interactions among clients are characterized by a time-varying, symmetric, and doubly stochastic mixing matrix $W^t \in \mathbb{R}^{n\times n}$, where $W_{ij}^t > 0$ if clients i and j can communicate in round t, and $W_{ij}^t = 0$ otherwise. W^t may vary across rounds.

Each client i draws training samples from a local data distribution $\mathcal{D}_i$, which may differ across clients. For a matrix parameter block $X \in \mathbb{R}^{m\times d}$, the collaborative optimization objective is

$$\min_X f(X) = \frac{1}{n}\sum_{i=1}^{n} f_i(X), \qquad f_i(X) = \mathbb{E}_{\xi\sim\mathcal{D}_i} F_i(X;\xi). \tag{1}$$

All clients are initialized from the same parameter X_0 with zero momentum. We use $r = \min\{m, d\}$ and denote the Frobenius, operator, and nuclear norms by $\|\cdot\|_F$, $\|\cdot\|_{\mathrm{op}}$, and $\|\cdot\|_*$, respectively.

3.2 ALGORITHM DESIGN

Given the limited memory and communication resources of decentralized clients, maintaining full-precision optimizer states during local training and exchanging full-precision model parameters with neighboring clients can incur substantial overhead. We therefore design a quantized decentralized Muon framework that supports a different feasible precision configuration for every client and communication round. The configurations may be supplied by a local resource monitor or an external scheduler; the framework does not assume a particular resource-allocation controller. It then uses blockwise stochastic rounding to reduce the storage cost of local training states, especially Muon's stored momentum state, and the communication payload exchanged among neighboring clients.

Definition 1 (Blockwise stochastic quantizer)**.** For an integer $b \geq 2$, let $L_b = 2^{b-1} - 1$. Partition a tensor V into a fixed collection of disjoint blocks $\{V^{(1)}, \dots, V^{(R)}\}$. For each block $V^{(r)} = [v_1^{(r)}, \dots, v_{d_r}^{(r)}]^\top \in \mathbb{R}^{d_r}$, all blocks share the same bit width b and integer quantization levels $\{-L_b, \dots, L_b\}$, while each block uses its own grid spacing

$$\delta_b(V^{(r)}) = \frac{\left\|V^{(r)}\right\|_\infty}{L_b}, \qquad r = 1, \dots, R, \tag{2}$$

where $\left\|V^{(r)}\right\|_\infty$ denotes the maximum absolute entry of block $V^{(r)}$. For a nonzero block $V^{(r)}$, define its quantized representation as

$$\mathcal{Q}(V^{(r)}; b) = \left[Q(v_1^{(r)}; b), \dots, Q(v_{d_r}^{(r)}; b)\right]^\top, \tag{3}$$

Algorithm 1: *RADeMuon*

Input: X_0, $\beta \in [0,1)$, K, T, stepsizes η_t, mixing matrices W^t, feasible precision configurations $(b^x_{i,t}, b^g_{i,t}, b^M_{i,t})$, blockwise stochastic quantizer $\mathcal{Q}(\cdot; b)$, NS degree κ, and iteration count J.

Initialize: $X_{i,0,0} = X_0$ and $M^Q_{i,0,0} = 0$ for every client i.

1 **for** $t = 0, \dots, T-1$ **do**

2 **for** *each client* i *in parallel* **do**

3 Set the feasible bit widths $(b^x_{i,t}, b^g_{i,t}, b^M_{i,t})$ supplied for the current client and round;

4 $X^Q_{i,t,0} \leftarrow \mathcal{Q}(X_{i,t,0}; b^x_{i,t})$;

5 **for** $k = 0, \dots, K-1$ **do**

6 $g^Q_{i,t,k} \leftarrow \mathcal{Q}\Big(\nabla F_i(X^Q_{i,t,k}; \xi_{i,t,k}); b^g_{i,t}\Big)$;

7 $M_{i,t,k+1} \leftarrow \beta M^Q_{i,t,k} + (1-\beta) g^Q_{i,t,k}$;

8 $M^Q_{i,t,k+1} \leftarrow \mathcal{Q}(M_{i,t,k+1}; b^M_{i,t})$; `// stored momentum state`

9 $O_{i,t,k} \leftarrow \mathrm{NS}_\kappa(M^Q_{i,t,k+1}; J)$; `// Algorithm 2`

10 $X^Q_{i,t,k+1} \leftarrow \mathcal{Q}\Big(X^Q_{i,t,k} - \eta_t O_{i,t,k}; b^x_{i,t}\Big)$; `// quantized parameter update`

11 Transmit the integer codes and block scales representing $X^Q_{i,t,K}$ to neighboring clients;

12 $X_{i,t+1,0} \leftarrow \sum_j W^t_{ij} X^Q_{j,t,K}$, forall i

13 $M^Q_{i,t+1,0} \leftarrow M^Q_{i,t,K}$ for all i ; `// local momentum carry`

where, for each entry $v^{(r)}_a$, let $p^{(r)}_a = v^{(r)}_a/\delta_b(V^{(r)}) - \lfloor v^{(r)}_a/\delta_b(V^{(r)})\rfloor$, and

$$Q(v^{(r)}_a; b) = \begin{cases} \delta_b(V^{(r)})\lfloor v^{(r)}_a/\delta_b(V^{(r)})\rfloor, & \text{w.p. } 1-p^{(r)}_a, \\ \delta_b(V^{(r)})\lceil v^{(r)}_a/\delta_b(V^{(r)})\rceil, & \text{w.p. } p^{(r)}_a. \end{cases} \tag{4}$$

The quantized tensor $\mathcal{Q}(V; b)$ is obtained by concatenating the quantized blocks $\{\mathcal{Q}(V^{(1)}; b), \dots, \mathcal{Q}(V^{(R)}; b)\}$. For a zero block, we set its scale to zero.

At the beginning of round t, client i receives a feasible configuration of three bit widths $(b^x_{i,t}, b^g_{i,t}, b^M_{i,t})$ for the model parameters, gradients, and momentum states, respectively (Line 3). These widths may vary across clients and rounds according to the supplied resource budgets, but remain fixed for all quantizer calls within the corresponding client-round. For each tensor, the selected width is shared by all blocks, while each block derives its own scale from (2). Client i first quantizes the round-start model parameter,

$$X^Q_{i,t,0} = \mathcal{Q}(X_{i,t,0}; b^x_{i,t}), \tag{5}$$

and performs local training on this quantized representation. At local step k, it evaluates a stochastic gradient at the current quantized parameter and quantizes the result using the selected gradient width,

$$g^Q_{i,t,k} = \mathcal{Q}\Big(\nabla F_i(X^Q_{i,t,k}; \xi_{i,t,k}); b^g_{i,t}\Big). \tag{6}$$

The width $b^g_{i,t}$ controls the precision of the gradient representation used during local training. The quantized gradient is then used to update the momentum, and the updated momentum is quantized using the selected momentum width,

$$M_{i,t,k+1} = \beta M^Q_{i,t,k} + (1-\beta) g^Q_{i,t,k}, \qquad M^Q_{i,t,k+1} = \mathcal{Q}(M_{i,t,k+1}; b^M_{i,t}). \tag{7}$$

The quantized momentum is stored as the optimizer state. At the beginning of a new round, the momentum state is carried over locally as $M^Q_{i,t,0} = M^Q_{i,t-1,K}$ for $t > 0$, with $M^Q_{i,0,0} = 0$. Hence, the newly selected $b^M_{i,t}$ is applied when the updated momentum is written during round t, without introducing an additional re-quantization at the round boundary.

The quantized momentum is subsequently orthogonalized through a finite Newton–Schulz (NS) iteration to obtain the local update direction,

$$O_{i,t,k} = \mathrm{NS}_\kappa(M^Q_{i,t,k+1}; J). \tag{8}$$

Rather than computing the polar factor through an exact SVD, the NS iteration approximates the orthogonalization direction using only matrix multiplications. Following the Newton–Schulz polynomial construction (Kim & Oh, 2026), for $M^Q_{i,t,k+1} \neq 0$, we initialize $U_0 = M^Q_{i,t,k+1}/\max\{1, \|M^Q_{i,t,k+1}\|_F\}$ and recursively apply

$$U_{j+1} = p_\kappa(U_j U_j^\top)U_j, \qquad p_\kappa(\lambda) = \sum_{a=0}^{\kappa} \frac{\binom{2a}{a}}{4^a}(1-\lambda)^a, \quad j = 0, \ldots, J-1. \tag{9}$$

Here, p_κ is the degree-κ Taylor approximation of $\lambda^{-1/2}$ around $\lambda = 1$. Repeatedly applying this polynomial drives $U_j U_j^\top$ toward the identity on the range of the momentum matrix, so that the final iterate U_J approaches its polar factor without an explicit SVD. We therefore set $O_{i,t,k} = U_J$. This SVD-free construction follows the practical Muon design, where a small number of NS steps is used to approximate momentum orthogonalization through matrix multiplications.

Algorithm 2: Finite NS map $\mathrm{NS}_\kappa(M^Q; J)$

Input: M^Q, κ, J; **output:** O;
if $M^Q = 0$ **then return** 0;
$U_0 \leftarrow M^Q/\max\{1, \|M^Q\|_F\}$;
for $j = 0, \ldots, J-1$ **do** $U_{j+1} \leftarrow p_\kappa(U_j U_j^\top)U_j$ using (9);
return $O \leftarrow U_J$;

The resulting direction is then applied to the current quantized parameter, and the updated parameter is directly written in the selected parameter format,

$$X^Q_{i,t,k+1} = \mathcal{Q}\Big(X^Q_{i,t,k} - \eta_t O_{i,t,k}; b^x_{i,t}\Big). \tag{10}$$

Consequently, the local parameter remains in the selected low-precision representation throughout the local updates.

After K local steps, the final representation $X^Q_{i,t,K}$ is already quantized. Client i therefore directly transmits its integer codes and associated block scales to its neighbors without an additional quantization call (Line 11). Hence, $b^x_{i,t}$ directly controls the communication cost of model exchange. Upon reception, the transmitted codes are reconstructed using their associated block scales, and the resulting numerical tensors are mixed as

$$X_{i,t+1,0} = \sum_j W^t_{ij} X^Q_{j,t,K}, \tag{11}$$

while the quantized momentum remains local and is carried to the next round (Lines 12–13).

Overall, $b^g_{i,t}$ controls the precision of the gradient representation during local training, $b^M_{i,t}$ controls the stored momentum state, and $b^x_{i,t}$ controls both the local parameter representation and the communication payload. The corresponding quantization errors, together with the approximation residual introduced by the finite NS iteration, are analyzed in Section 4.

4 Convergence Analysis

We next characterize how heterogeneous and time-varying quantization, decentralized communication, and finite Newton–Schulz (NS) orthogonalization jointly affect the convergence of *RADeMuon*. Rather than reproducing the technical recursions used in the proof, we retain only the key factors that determine the convergence behavior. The complete derivation is deferred to Appendix A.

4.1 Assumptions and basic guarantees

Assumption 1 (Smoothness and lower boundedness)**.** The global objective is lower bounded by $f^\star := \inf_X f(X) > -\infty$. Each local objective f_i is continuously differentiable and L-smooth with respect to the spectral norm and its dual nuclear norm, i.e., for all X, Y,

$$\|\nabla f_i(X) - \nabla f_i(Y)\|_* \leq L\,\|X - Y\|_{\mathrm{op}}. \tag{12}$$

Assumption 2 (Bounded stochastic variance)**.** For every client i, the stochastic gradient is an unbiased estimator of the local gradient and has bounded variance,

$$\mathbb{E}[\nabla F_i(X;\xi)] = \nabla f_i(X), \qquad \mathbb{E}\Big[\|\nabla F_i(X;\xi) - \nabla f_i(X)\|_F^2\Big] \le \sigma_1^2. \tag{13}$$

Assumption 3 (Bounded data heterogeneity)**.** The gradient discrepancy among clients is bounded as

$$\|\nabla f_i(X) - \nabla f(X)\|_F^2 \le \sigma_2^2, \qquad \forall X, \tag{14}$$

Assumption 4 (Network mixing)**.** At round t, the mixing matrix $W^t = [W_{ij}^t] \in \mathbb{R}^{n\times n}$ is nonnegative, symmetric, and doubly stochastic, satisfying

$$W^t\mathbf{1} = \mathbf{1}, \qquad \mathbf{1}^\top W^t = \mathbf{1}^\top, \qquad W^t = (W^t)^\top, \qquad 0 \le W_{ij}^t \le 1. \tag{15}$$

Let $J_n = \mathbf{1}\mathbf{1}^\top/n$ and define the contraction factor

$$\rho_t := (\max\{|\lambda_2(W^t)|, |\lambda_n(W^t)|\})^2 = \left\|W^t - J_n\right\|_{\mathrm{op}}^2, \qquad \rho := \max_{0\le t<T} \rho_t < 1. \tag{16}$$

Lemma 1 (Geometric contraction of mixing products)**.** *Under Assumption 4, for any $0 \le u \le t \le T$, we have*

$$\left\|(I - J_n)\prod_{\ell=u}^{t-1} W^\ell\right\|_{\mathrm{op}} \le \rho^{(t-u)/2}. \tag{17}$$

Lemma 2 (Blockwise quantization guarantee)**.** *Consider the blockwise stochastic quantizer in Definition 1. For a selected bit width b and maximum block size s, the quantizer is unbiased and satisfies*

$$\mathbb{E}[\mathcal{Q}(V;b)] = V, \qquad \mathbb{E}[\|\mathcal{Q}(V;b) - V\|_F^2] \le q(b)\,\|V\|_F^2, \tag{18}$$

where $q(b) = \frac{s}{4(2^{b-1}-1)^2}$. Accordingly, the three precisions specified for client i in round t induce

$$q_{i,t}^x = q(b_{i,t}^x), \qquad q_{i,t}^g = q(b_{i,t}^g), \qquad q_{i,t}^M = q(b_{i,t}^M), \tag{19}$$

for the parameter, gradient, and momentum quantization, respectively.

Lemma 3 (Finite Newton–Schulz approximation)**.** *For the Newton–Schulz iteration in Algorithm 2, let $U_{i,t,k,0}$ denote the initialization associated with $M_{i,t,k+1}^Q$, and let $\Pi_{i,t,k}$ be the orthogonal projector onto the column space of $M_{i,t,k+1}^Q$. Define the initial orthogonality residual as $\delta_{i,t,k,0} := \|\Pi_{i,t,k} - U_{i,t,k,0}U_{i,t,k,0}^\top\|_{\mathrm{op}}$, and let $\delta_0 := \sup_{i,t,k}\delta_{i,t,k,0} < 1$. After J iterations of the degree-κ Newton–Schulz polynomial,*

$$\left\|\mathrm{NS}_\kappa(M_{i,t,k+1}^Q; J) - \mathrm{Polar}(M_{i,t,k+1}^Q)\right\|_{\mathrm{op}} \le \varepsilon_J, \qquad \varepsilon_J := \delta_0^{(\kappa+1)^J} < 1. \tag{20}$$

Consequently,

$$\|O_{i,t,k}\|_F \le \sqrt{r}(1+\varepsilon_J), \qquad r := \min\{m,d\}. \tag{21}$$

The quantity ε_J characterizes the finite-orthogonalization error. It decreases rapidly with the NS iteration count J and polynomial degree κ, while the quantization of the momentum changes the input on which this approximation is performed.

4.2 Main convergence result

Theorem 1 (Convergence of *RADeMuon*)**.** *Suppose Assumptions 1–4 hold. Let $0 \le \beta < 1$, and let $\{\eta_t\}_{t=0}^{T-1}$ be positive and non-increasing, with $\bar\eta \triangleq T^{-1}\sum_{t=0}^{T-1}\eta_t$. Let $\Delta \in (0,1]$ be the positive descent-margin constant defined explicitly in Appendix* (S.57)*. Then, the iterates satisfy*

$$\begin{aligned}
\frac{1}{T\bar\eta}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)\right\|_* &\le \frac{\mathbb{E}\left[f\left(\mathbf{X}_0\frac{\mathbf{1}}{n}\right)\right] - f^\star}{\Delta KT\bar\eta(1-\varepsilon_J)} + \left[\frac{6\sqrt{r}}{(1-\varepsilon_J)\Delta} + \frac{1-\Delta}{\Delta}\right](\sigma_1+\sigma_2) \\
&+ \left\{\frac{LrK}{(1-\varepsilon_J)\Delta}\left[(1+\varepsilon_J)^2 + \frac{8(1+\varepsilon_J)}{(1-\beta)(1-\sqrt{\rho})}\right] + \frac{10LKr\sqrt{n}(1+\varepsilon_J)}{(1-\varepsilon_J)(1-\sqrt{\rho})\Delta}\right\}\frac{\sum_{t=0}^{T-1}\eta_t^2}{T\bar\eta} \\
&+ \left[\frac{6L\sqrt{r}}{\sqrt{n}(1-\varepsilon_J)(1-\beta)(1-\sqrt{\rho})\Delta} + \frac{10L\sqrt{r}}{(1-\varepsilon_J)(1-\sqrt{\rho})\Delta}\right]\frac{1}{T\bar\eta}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^x}\,\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F \\
&+ \frac{L(K+1)}{\Delta nK(1-\varepsilon_J)T\bar\eta}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\sum_{k=0}^{K} q_{i,t}^x\mathbb{E}\|\mathbf{x}_{i,t,k}\|_F^2.
\end{aligned} \tag{22}$$

Remark 1. Theorem 1 establishes a convergence guarantee for quantized decentralized Muon with finite-step Newton–Schulz orthogonalization. To the best of our knowledge, this is the first convergence result for distributed Muon that simultaneously accounts for the coupled algorithmic and system-induced errors arising in practical decentralized training. The decomposition of the convergence bound highlights three critical factors:

- **Network Topology:** The effect of decentralized communication is characterized by the spectral factor $(1-\sqrt{\rho})^{-1}$. A smaller ρ corresponds to better network connectivity and yields a tighter convergence bound by reducing the accumulation of model disagreement across clients.
- **Newton–Schulz Approximation:** The error introduced by finite-step Newton–Schulz orthogonalization is captured by the factor $(1-\varepsilon_J)^{-1}$. As the approximation error ε_J decreases, the convergence guarantee becomes tighter and progressively approaches that of the exact SVD-based polar update.
- **Quantization:** The impact of quantization is characterized by the terms involving $\sqrt{q_{i,t}^x}\,\mathbb{E}\|\mathbf{x}_{i,t,j}\|_{\mathrm{F}}$ and $q_{i,t}^x\,\mathbb{E}\|\mathbf{x}_{i,t,k}\|_{\mathrm{F}}^2$. More aggressive quantization introduces larger optimization errors, whereas finer quantization yields a tighter convergence bound. Moreover, the former error is further amplified by $(1-\sqrt{\rho})^{-1}$, revealing the coupling between quantization distortion and decentralized model disagreement.

The gradient and momentum quantization errors $q_{i,t}^g$ and $q_{i,t}^M$ enter through the descent margin Δ, whose full expression is given in Appendix (S.57). Therefore, the theorem provides a unified characterization of how quantization, network connectivity, and approximate orthogonalization jointly affect the convergence of distributed Muon.

Corollary 1. *Under the conditions of Theorem 1, choose* $\eta_t = D_t\sqrt{\frac{1-\beta}{KT}}$, $D_m = \min_{0\le t<T} D_t$, $D_M = \max_{0\le t<T} D_t$, *where* $\{D_t\}_{t=0}^{T-1}$ *is a positive and non-increasing sequence of round-dependent learning-rate scaling factors, i.e.,* $D_{t+1} \le D_t$, *and* $\beta = 1 - \min\left\{\sqrt{\frac{nL}{rKT}}, 1\right\}$. *Then, the convergence bound can be summarized as*

$$
\begin{aligned}
\frac{1}{T\bar{\eta}}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)\right\|_* &\le \frac{r^{1/4}(\mathbb{E}\left[f\left(\mathbf{X}_0\frac{\mathbf{1}}{n}\right)\right]-f^\star)}{\Delta D_m(1-\varepsilon_J)(nLKT)^{1/4}} \\
&+\left[\frac{6\sqrt{r}}{(1-\varepsilon_J)\Delta}+\frac{1-\Delta}{\Delta}\right](\sigma_1+\sigma_2)+\frac{D_M}{(1-\varepsilon_J)\Delta}\left(\frac{L^5r^3Kn}{T^3}\right)^{\frac{1}{4}}\left[(1+\varepsilon_J)^2+\frac{1+\varepsilon_J}{1-\sqrt{\rho}}\left(8\sqrt{\frac{rKT}{nL}}+10\sqrt{n}\right)\right] \\
&+\frac{L\sqrt{r}\,D_M}{D_m(1-\varepsilon_J)(1-\sqrt{\rho})\Delta}\left[10+6\sqrt{\frac{rKT}{n^2L}}\right]\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^x}\,\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F \\
&+\frac{L^{3/4}r^{1/4}(K+1)K^{-1/4}}{\Delta D_m(1-\varepsilon_J)n^{5/4}T^{1/4}}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\sum_{k=0}^{K}q_{i,t}^x\mathbb{E}\|\mathbf{x}_{i,t,k}\|_F^2. \qquad (23)
\end{aligned}
$$

Remark 2 (Convergence rate). Corollary 1 shows that the convergence bound contains a dominant $\mathcal{O}\big((nKT)^{-1/4}\big)$ stationarity term when the quantization-induced terms are negligible. Convergence to a first-order stationary point further requires the stochastic and heterogeneity residuals to vanish. The factor $n^{-1/4}$ describes the client dependence of the dominant term when Δ, the network factor, and the residual terms remain controlled.

5 Experiments

5.1 Setup

We evaluate image classification using ResNet-18 on TinyImageNet-200 (Le et al., 2015) and language-model pretraining using a 124M nanoGPT model (Karpathy, 2023) on FineWeb-Edu (Penedo et al., 2024). For image classification, the data are partitioned non-IID across clients using a Dirichlet distribution with $\alpha \in \{0.1, 0.5\}$, and training lasts 300 communication rounds

Table 2: ResNet-18 on TinyImageNet-200 under two label-skew settings. Acc@1/Acc@5 denote top-1/top-5 accuracy; Comp. and Comm. follow Section 5.

Algorithm	Bit	$\mathrm{Dir}(\alpha = 0.1)$		$\mathrm{Dir}(\alpha = 0.5)$		Efficiency	
		Acc@1 (%)	Acc@5 (%)	Acc@1 (%)	Acc@5 (%)	Comp. (MB)	Comm. (MB)
DPSGD	32	54.80	78.31	57.21	79.08	43.00	494.47
LM-DFL	32	54.56	78.08	57.07	79.61	43.00	988.93
	16	55.58	78.12	57.15	79.70	43.00	494.47
	8	55.16	78.10	57.02	79.21	43.00	247.23
LPT	32	55.07	78.06	57.28	79.64	43.00	494.47
	16	55.04	78.35	57.37	79.38	21.50	247.23
	8	51.61	74.99	56.00	78.95	10.75	123.62
Adriana	32	55.26	78.22	57.02	79.74	43.00	186.08
	16	55.79	78.67	56.61	79.52	21.50	186.00
	8	55.07	78.16	57.13	79.46	10.75	185.08
De-EF21-Muon	32	59.11	80.12	62.25	82.06	43.00	494.47
	16	59.01	80.50	61.37	82.04	43.00	247.67
	8	59.00	80.32	61.24	81.85	43.00	124.16
RADeMuon	–	58.96	80.44	61.91	81.93	17.92	204.38

with 50 local steps per round. For language-model pretraining, each client performs 30 local steps per round under a total training budget of 2.5B tokens. We additionally evaluate parameter-efficient fine-tuning of GPT-2 XL with LoRA on seven GLUE tasks (Wang et al., 2018), with the corresponding setup and results reported in Appendix B. In all experiments, five clients communicate through a randomly generated doubly stochastic matrix constructed using Metropolis–Hastings weights (Awan et al., 2006).

We compare with DPSGD (Lian et al., 2017), LM-DFL (Chen et al., 2024), LPT (Li et al., 2024), Adriana (Qiao et al., 2026), and De-EF21-Muon (Gruntkowska et al., 2026). To model changing resource budgets, *RADeMuon* uses 32 bits in the first $T/6$ rounds, 16 bits in the next $T/6$, and 8 bits in the remaining $2T/3$ for all three representations. This externally supplied profile is a temporal special case of our framework; baseline rows at 32, 16, or 8 bits are separate fixed-precision runs. Quantized tensors use blocks of at most 256 entries. Following Adriana, both resource metrics are reported in MB. Comp. denotes the average local computation cost per client, whereas Comm. denotes the total communication volume of all five clients over all communication rounds and accounts for the topology used in each round. Full settings are in Appendix B.

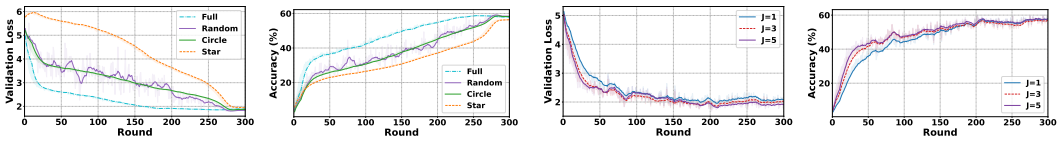
5.2 Evaluation

Impact of Resource-Aware Quantization. Tables 2 and 3 report the results on TinyImageNet-200 and FineWeb-Edu, respectively. On TinyImageNet-200, *RADeMuon* achieves Acc@1 of 58.96% and 61.91% under $\alpha = 0.1$ and $\alpha = 0.5$, respectively, only 0.15 and 0.34 percentage points below 32-bit De-EF21-Muon. At the same time, it reduces local computation and communication costs by 58.33% and 58.67% relative to that baseline. Compared with full-precision DPSGD, *RADeMuon* improves Acc@1 by 4.16 and 4.70 percentage points. It also outperforms the strongest non-Muon baseline by 3.17 and 4.54 percentage points under the two partitions. These results show that varying the precision with the resource budget preserves the optimization advantage of Muon under different levels of data heterogeneity while substantially reducing resource costs.

For nanoGPT pretraining on FineWeb-Edu, *RADeMuon* achieves a validation perplexity of 24.51 and an average zero-shot accuracy of 38.58% across the 12 downstream tasks. Compared with 32-bit De-EF21-Muon, it reduces local computation and communication costs by 58.34% and 58.42%, respectively. Its perplexity improves from 24.57 to 24.51, while its average zero-shot accuracy decreases by only 0.36 percentage points, from 38.94% to 38.58%. Moreover, *RADeMuon* exceeds 32-bit LPT, Adriana, and the strongest non-Muon baseline by 2.06, 1.71, and 1.27 percentage points in average downstream accuracy, respectively. Fixed 8-bit LPT and Adriana exhibit clear degradation in perplexity or downstream accuracy, whereas the resource-aware precision configuration retains the pretraining and downstream quality of full-precision De-EF21-Muon while substantially reducing its resource costs. Together, the two tasks show that *RADeMuon* maintains the benefit of Muon while accommodating varying resource budgets.

Table 3: NanoGPT (124M) pretraining on FineWeb-Edu and zero-shot evaluation. Val. CE and PPL denote validation cross-entropy and perplexity; Comp. and Comm. follow Section 5.

Algorithm	Bit	Pretraining		Downstream tasks evaluation					Efficiency	
		Val. CE	PPL	HellaSwag	ARC-c	ARC-e	PIQA	Avg. 12 tasks	Comp. (MB)	Comm. (MB)
DPSGD	32	3.45	31.62	29.58	24.06	42.47	58.65	36.81	471.45	5375.67
LM-DFL	32	3.46	31.70	29.28	25.34	42.63	59.47	37.31	471.45	10751.34
	16	3.46	31.66	29.45	24.23	42.47	59.09	36.79	471.45	5636.75
	8	3.46	31.97	29.36	24.15	42.55	58.76	36.84	471.45	2687.83
LPT	32	3.46	31.84	29.28	25.00	41.33	59.41	36.52	471.45	5375.67
	16	3.46	31.81	28.96	24.83	41.58	58.22	36.98	235.72	2687.84
	8	3.49	32.83	27.96	22.87	38.34	57.18	35.18	117.86	1343.92
Adriana	32	3.45	31.48	29.45	25.43	42.47	59.41	36.87	471.45	2279.27
	16	3.45	31.64	29.66	23.46	41.88	58.16	36.37	235.72	2223.06
	8	3.69	40.40	27.30	24.06	37.71	56.31	34.73	117.86	2199.61
De-EF21-Muon	32	3.20	24.57	30.27	25.94	46.04	60.39	38.94	471.45	5375.67
	16	3.19	24.32	30.44	24.15	45.58	59.96	38.18	471.45	2694.05
	8	3.19	24.29	30.46	26.28	45.08	60.50	38.10	471.45	1416.03
RADeMuon	–	3.20	24.51	29.80	25.09	44.99	60.45	38.58	196.40	2235.38



(a) Communication topology (b) Number of NS iterations

Figure 1: Ablation studies on TinyImageNet-200 with $\mathrm{Dir}(\alpha = 0.1)$. Panel (a) reports the effect of communication topology on validation cross-entropy and Acc@1. Panel (b) reports the effect of the number of Newton–Schulz iterations.

Impact of Network Topology. We train ResNet-18 on TinyImageNet-200 with $\mathrm{Dir}(\alpha = 0.1)$ under fully connected, random, circle, and star topologies. Fully connected communication links every pair of clients, the circle uses two immediate neighbors, the star relays information through a central client, and the random graph follows the main experimental setting. Figure 1(a) shows that the fully connected topology converges fastest. This agrees with our analysis: better connectivity yields a smaller contraction factor ρ and reduces $(1-\sqrt{\rho})^{-1}$, thereby limiting inter-client disagreement and the amplification of quantization errors.

Impact of Newton–Schulz Iterations. We vary the number of NS iterations over $J \in \{1, 3, 5\}$ while keeping the remaining settings fixed. Figure 1(b) shows that increasing J generally accelerates convergence over the evaluated range, whereas fewer iterations lead to slower optimization progress. This behavior is consistent with our finite-NS analysis: a larger J reduces the approximation error ε_J, brings the update direction closer to the exact polar direction, and decreases the factor $(1-\varepsilon_J)^{-1}$. Additional NS iterations therefore tighten the finite-iteration guarantee without changing the asymptotic order, but they require more computation at every local step.

6 Conclusion

We proposed *RADeMuon*, a resource-aware quantized framework for fully decentralized Muon that supports precision configurations following the available resources throughout training. It quantizes parameters, gradients, and momentum states and communicates the resulting parameter buffer directly. Our finite-horizon analysis has a leading $\mathcal{O}((nKT)^{-1/4})$ stationarity term under a positive descent margin and vanishing residual terms, while accounting for momentum accumulation, finite-step NS, and decentralized aggregation. Across image classification and language-model pretraining, *RADeMuon* reduces local computation and communication costs by up to 58.34% and 58.67%, respectively, relative to full-precision De-EF21-Muon while maintaining comparable model quality.

AI USE STATEMENT

Generative AI tools were used for language editing, structural revision, and literature cross-checking. No experimental values were generated by these tools. The authors reviewed the final text and take responsibility for its content.

REPRODUCIBILITY STATEMENT

We have made efforts to support the reproducibility of both the proposed method and the reported experimental results. The main paper describes the optimization objective, quantizer, local updates, finite-step NS procedure, and decentralized aggregation. The appendix provides the assumptions and complete proof details. The experimental section and appendix describe the datasets, model architectures, baselines, evaluation protocols, and training configurations used in our experiments.

REFERENCES

Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed Orthonormalized Updates. *arXiv preprint arXiv:2504.05295*, 2025.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1709–1720, 2017.

Asad Awan, Ronaldo A Ferreira, Suresh Jagannathan, and Ananth Grama. Distributed uniform sampling in unstructured peer-to-peer networks. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 9, pp. 223c–223c. IEEE, 2006.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Li Chen, Wei Liu, Yunfei Chen, and Weidong Wang. Communication-efficient design for quantized decentralized federated learning. *IEEE Transactions on Signal Processing*, 72:1175–1188, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pp. 177–190. Springer, 2005.

William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the third international workshop on paraphrasing (IWP2005)*, 2005.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Zenodo*, 2021.

Kaja Gruntkowska, Alexander Gaponov, Zhirayr Tovmasyan, and Peter Richtárik. Error Feedback for Muon and Friends. In *International Conference on Learning Representations*, 2026.

Aman Gupta, Rafael Celente, Abhishek Shivanna, D. T. Braithwaite, Gregory Dexter, Shao Tang, Hiroto Udagawa, Daniel Silva, Rohan Ramanath, and S. Sathiya Keerthi. Effective Quantization of Muon Optimizer States. *arXiv preprint arXiv:2509.23106*, 2025.

Chuan He, Shuyi Ren, Jingwei Mao, and Erik G. Larsson. DeMuon: A Decentralized Muon for Matrix Optimization over Graphs. *arXiv preprint arXiv:2510.01377*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error Feedback Fixes SignSGD and Other Gradient Compression Schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261. PMLR, 2019.

Andrej Karpathy. nanogpt. GitHub repository, 2023.

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3478–3487. PMLR, 2019.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Zhiwei Li, Yiqiu Li, Binbin Lin, Zhongming Jin, and Weizhong Zhang. Low precision local training is enough for federated learning. *Advances in Neural Information Processing Systems*, 37:90160–90197, 2024.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is Scalable for LLM Training. *arXiv preprint arXiv:2502.16982*, 2025.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1525–1534, 2016.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Jing Qiao, Yu Liu, Yuan Yuan, Yifei Zou, Xiao Zhang, and Dongxiao Yu. Resource-Aware Decentralized Learning with Rate-Adaptive Quantization. In *IEEE INFOCOM 2026 – IEEE Conference on Computer Communications*, pp. 1–10. IEEE, 2026. doi: 10.1109/INFOCOM59046.2026.11571464.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95, 2011.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Yupeng Su, Ruijie Zhang, Ziyue Liu, Yequan Zhao, and Zheng Zhang. MuonQ: Enhancing low-bit muon quantization via directional fidelity optimization. In *Conference on Language Modeling*, 2026.

Yuki Takezawa, Anastasia Koloskova, Xiaowen Jiang, and Sebastian U. Stich. FedMuon: Federated Learning with Bias-Corrected LMO-Based Optimization. *arXiv preprint arXiv:2509.26337*, 2025.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical Low-Rank Gradient Compression for Distributed Optimization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14236–14245, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pp. 1112–1122, 2018.

Huaijin Wu, Bingrui Li, Yebin Yang, Yi Tu, Zhanpeng Zhou, Jianfei Chen, and Junchi Yan. Achieving low-bit muon through subspace preservation and grid quantization. In *International Conference on Learning Representations*, 2026.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Hengrui Zhang, Boao Kong, Jiahe Geng, and Zhengyang Huang. SUDA-Muon: Structural Design Principles and Boundaries for Fully Decentralized Muon. *arXiv preprint arXiv:2604.23980*, 2026.

A PROOF OF THEOREM 1

A.1 PRELIMINARIES AND TECHNICAL LEMMAS

Notation and Subscript Convention. Throughout the convergence proof, we distinguish single-, double-, and triple-subscript quantities according to the following system-wide convention. Empty sums, empty products, and zeroth powers are zero, one, and one.

- **Three subscripts** (i,t,k) **(Local Client Quantities):** Variables with three subscripts represent local $m \times d_{\mathrm{m}}$ matrix states at client $i \in \{1,\dots,n\}$, global communication round $t \in \{0,\dots,T-1\}$, and local step $k \in \{0,\dots,K\}$. Specifically, $\mathbf{x}_{i,t,k}$ denotes the local weight iterate (including the pre-communication endpoint $k=K$), $\mathbf{O}_{i,t,k} \triangleq \mathrm{NS}_\kappa(\mathbf{M}^Q_{i,t,k+1}; J)$ is defined for $k \in \{0,\dots,K-1\}$, and $\mathbf{e}^x_{i,t,k} \triangleq \mathbf{x}^Q_{i,t,k} - \mathbf{x}_{i,t,k}$ denotes the local forward weight quantization error.
- **Two subscripts** (t,k) **(Global Concatenated Quantities):** Variables with two subscripts represent the network-wide horizontal concatenation of all n local $m \times d_{\mathrm{m}}$ matrices into an $m \times (nd_{\mathrm{m}})$ block matrix at round t and step k. Specifically, $\mathbf{O}_{t,k} \triangleq [\mathbf{O}_{1,t,k},\dots,\mathbf{O}_{n,t,k}] \in \mathbb{R}^{m\times(nd_{\mathrm{m}})}$ and $\mathbf{e}^x_{t,k} \triangleq [\mathbf{e}^x_{1,t,k},\dots,\mathbf{e}^x_{n,t,k}] \in \mathbb{R}^{m\times(nd_{\mathrm{m}})}$.
- **Single subscript** t **(Round-Initial Global Quantities):** $\mathbf{X}_t \triangleq [\mathbf{x}_{1,t,0},\dots,\mathbf{x}_{n,t,0}] \in \mathbb{R}^{m\times(nd_{\mathrm{m}})}$ denotes the concatenated global model parameter matrix at the beginning of round t (i.e., at local step $k=0$).

Let $\mathbf{I} \in \mathbb{R}^{n\times n}$ denote the identity matrix, and let $\mathbf{J} \triangleq \frac{\mathbf{1}\mathbf{1}^\top}{n} \in \mathbb{R}^{n\times n}$ denote the uniform network averaging projection matrix. Throughout this paper, right multiplication of any concatenated state $\mathbf{A} \in \mathbb{R}^{m\times(nd_{\mathrm{m}})}$ by an $n \times n$ network matrix (e.g., $\mathbf{J}$ or $\mathbf{W}^t$) is implicitly defined block-wise over the n column blocks of size $m \times d_{\mathrm{m}}$. Consequently, $\mathbf{X}_t\mathbf{J} = [\mathbf{X}_t\frac{\mathbf{1}}{n},\dots,\mathbf{X}_t\frac{\mathbf{1}}{n}]$ represents the global network-averaged state, and $\mathbf{X}_t(\mathbf{I}-\mathbf{J})$ characterizes the network consensus drift across all clients. $\mathcal{Q}^c_{i,t}(\cdot) \equiv Q(\cdot; b^c_{i,t})$ is the quantizer of Definition 1 ($c \in \{x,g,M\}$).

In Decentralized Quantized Muon (Algorithm 1), each client $i \in \{1,\dots,n\}$ updates its local model parameter at step $k \in \{0,\dots,K-1\}$ via $\mathbf{x}_{i,t,k+1} = \mathbf{x}^Q_{i,t,k} - \eta_t\mathbf{O}_{i,t,k} = \mathbf{x}_{i,t,k} + \mathbf{e}^x_{i,t,k} - \eta_t\mathbf{O}_{i,t,k}$, with local weight quantization error defined as $\mathbf{e}^x_{i,t,k} \triangleq \mathbf{x}^Q_{i,t,k} - \mathbf{x}_{i,t,k}$. The local momentum buffer evolves recursively as:

$$\mathbf{M}_{i,t,k+1} = \beta\mathbf{M}^Q_{i,t,k} + (1-\beta)\nabla^Q F_i(\mathbf{x}^Q_{i,t,k}; \xi_{i,t,k}), \tag{S.24}$$

where $\mathbf{M}^Q_{i,t,k} \triangleq \mathcal{Q}^M_{i,t}(\mathbf{M}_{i,t,k})$, with the unquantized buffer carried across rounds by the proof-side convention $\mathbf{M}_{i,t,0} = \mathbf{M}_{i,t-1,K}$ (Algorithm 1 carries only the quantized buffer $\mathbf{M}^Q$), and $\mathbf{M}^Q_{i,t,0} = \mathbf{M}^Q_{i,t-1,K}$ for $t \geq 1$, initialized at $\mathbf{M}_{i,0,0} = \mathbf{M}^Q_{i,0,0} = \mathbf{0}$. The update direction $\mathbf{O}_{i,t,k} \triangleq \mathrm{NS}_\kappa(\mathbf{M}^Q_{i,t,k+1}; J)$, i.e. the output $\mathbf{U}_J$ of the finite Newton–Schulz orthogonalization of Algorithm 2, approximates the polar factor $\mathbf{P}_{i,t,k} \triangleq \mathrm{Polar}(\mathbf{M}^Q_{i,t,k+1})$.

To decouple the multi-source errors arising from stochastic gradient noise, data heterogeneity, gradient quantization, forward weight quantization, and momentum buffer quantization, we define the following four auxiliary sequences initialized at $\mathbf{C}_{i,0,0} = \mathbf{S}_{i,0,0} = \mathbf{Y}_{i,0,0} = \mathbf{Z}_{i,0,0} = \mathbf{0}$:

$$\mathbf{C}_{i,t,k+1} = \beta\mathbf{C}_{i,t,k} + (1-\beta)\nabla f_i(\mathbf{x}_{i,t,k}), \qquad \mathbf{C}_{i,t,0} = \mathbf{C}_{i,t-1,K}, \tag{S.25}$$

$$\mathbf{S}_{i,t,k+1} = \beta\mathbf{S}_{i,t,k} + (1-\beta)\nabla F_i(\mathbf{x}_{i,t,k}; \xi_{i,t,k}), \qquad \mathbf{S}_{i,t,0} = \mathbf{S}_{i,t-1,K}, \tag{S.26}$$

$$\mathbf{Y}_{i,t,k+1} = \beta\mathbf{Y}_{i,t,k} + (1-\beta)\nabla^Q F_i(\mathbf{x}_{i,t,k}; \xi_{i,t,k}), \qquad \mathbf{Y}_{i,t,0} = \mathbf{Y}_{i,t-1,K}, \tag{S.27}$$

$$\mathbf{Z}_{i,t,k+1} = \beta\mathbf{Z}_{i,t,k} + (1-\beta)\nabla^Q F_i(\mathbf{x}^Q_{i,t,k}; \xi_{i,t,k}), \qquad \mathbf{Z}_{i,t,0} = \mathbf{Z}_{i,t-1,K}. \tag{S.28}$$

The auxiliary sequences are initialized with zero momentum and progressively introduce stochastic gradients, gradient quantization, and forward weight quantization. In particular, $\mathbf{C}$, $\mathbf{S}$, $\mathbf{Y}$, and $\mathbf{Z}$ correspond to the true-gradient, stochastic-gradient, gradient-quantized, and forward-quantized updates, respectively. Thus, the differences $\mathbf{C}-\mathbf{S}$, $\mathbf{S}-\mathbf{Y}$, $\mathbf{Y}-\mathbf{Z}$, and $\mathbf{Z}-\mathbf{M}^Q$ isolate the corresponding error sources.

Proof of Lemma 1.

Proof. Since every $\mathbf{W}^s$ is symmetric and doubly stochastic, $\mathbf{J}\mathbf{W}^s = \mathbf{W}^s\mathbf{J} = \mathbf{J}$. Hence,

$$\begin{aligned}
\left\|(\mathbf{I}-\mathbf{J})\prod_{s=\ell}^{t-1}\mathbf{W}^s\right\|_{\mathrm{op}} &= \left\|\prod_{s=\ell}^{t-1}\mathbf{W}^s - \mathbf{J}\right\|_{\mathrm{op}} \\
&= \left\|(\mathbf{I}-\mathbf{J})\prod_{s=\ell}^{t-1}(\mathbf{W}^s-\mathbf{J})\right\|_{\mathrm{op}} \\
&\le \|\mathbf{I}-\mathbf{J}\|_{\mathrm{op}}\prod_{s=\ell}^{t-1}\|\mathbf{W}^s-\mathbf{J}\|_{\mathrm{op}} \\
&= \prod_{s=\ell}^{t-1}\rho_s^{1/2} = \left(\prod_{s=\ell}^{t-1}\rho_s\right)^{1/2} \le \rho^{(t-\ell)/2}.
\end{aligned}$$

□

Lemma S.1 (Average Iterate Evolution across K Local Steps)**.** *Under Assumption 4, the network-averaged iterate across round $t \in \{0,1,\dots,T-1\}$ satisfies*

$$(\mathbf{X}_{t+1}-\mathbf{X}_t)\frac{\mathbf{1}}{n} = \sum_{k=0}^{K}\mathbf{e}_{t,k}^x\frac{\mathbf{1}}{n} - \eta_t\sum_{k=0}^{K-1}\mathbf{O}_{t,k}\frac{\mathbf{1}}{n}. \tag{S.29}$$

Proof. Telescoping the local updates gives the pre-communication state $\mathbf{X}_{t+1/2} = \mathbf{X}_t + \sum_{k=0}^{K}\mathbf{e}_{t,k}^x - \eta_t\sum_{k=0}^{K-1}\mathbf{O}_{t,k}$, where the term $\mathbf{e}_{t,K}^x$ is the forward-quantization error of the exchanged endpoints $\mathbf{x}_{i,t,K}^Q = \mathbf{x}_{i,t,K} + \mathbf{e}_{i,t,K}^x$. After communication, $\mathbf{X}_{t+1} = \mathbf{X}_{t+1/2}\mathbf{W}^t$. Right-multiplication by $\mathbf{1}/n$ and Assumption 4 give

$$\begin{aligned}
\mathbf{X}_{t+1}\frac{\mathbf{1}}{n} &= \left(\mathbf{X}_{t+1/2}\mathbf{W}^t\right)\frac{\mathbf{1}}{n} = \mathbf{X}_{t+1/2}\left(\mathbf{W}^t\frac{\mathbf{1}}{n}\right) = \mathbf{X}_{t+1/2}\frac{\mathbf{1}}{n} \\
&= \mathbf{X}_t\frac{\mathbf{1}}{n} + \sum_{k=0}^{K}\mathbf{e}_{t,k}^x\frac{\mathbf{1}}{n} - \eta_t\sum_{k=0}^{K-1}\mathbf{O}_{t,k}\frac{\mathbf{1}}{n}.
\end{aligned} \tag{S.30}$$

□

Lemma S.2 (Bound on the Update Direction)**.** *Under Lemma 3 of the main text, for any $i \in \{1,\dots,n\}$, round $t \ge 0$, and local step $k \in \{0,\dots,K-1\}$, the following bound holds:*

$$\left\|\sum_{k=0}^{K-1}\mathbf{O}_{t,k}\frac{\mathbf{1}}{n}\right\|_F^2 \le K^2 r(1+\varepsilon_J)^2,$$

where $r \triangleq \min\{m, d_{\mathrm{m}}\}$.

Proof. For the aggregate round direction, applying Cauchy–Schwarz and Jensen's inequality yields

$$\begin{aligned}
\left\|\sum_{k=0}^{K-1}\mathbf{O}_{t,k}\frac{\mathbf{1}}{n}\right\|_F^2 &\le K\sum_{k=0}^{K-1}\left\|\mathbf{O}_{t,k}\frac{\mathbf{1}}{n}\right\|_F^2 \\
&= K\sum_{k=0}^{K-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{O}_{i,t,k}\right\|_F^2 \\
&\le \frac{K}{n}\sum_{k=0}^{K-1}\sum_{i=1}^{n}\|\mathbf{O}_{i,t,k}\|_F^2 \\
&\le K^2 r(1+\varepsilon_J)^2.
\end{aligned} \tag{S.31}$$

□

Lemma S.3 (Bound on the Network Consensus Deviation)**.** *Under Assumptions 1–4, Lemmas 2 and 3 of the main text, and Lemma 1 of the main text, with positive stepsizes and* $\mathbf{X}_0(\mathbf{I}-\mathbf{J})=\mathbf{0}$*, for every* $0 \le t \le T$*,*

$$\mathbb{E}\,\|\mathbf{X}_t(\mathbf{I}-\mathbf{J})\|_F \le \sum_{\ell=0}^{t-1} \rho^{(t-\ell)/2}\left[\sum_{j=0}^{K}\sum_{i=1}^{n}\sqrt{q_{i,\ell}^x}\,\mathbb{E}\,\|\mathbf{x}_{i,\ell,j}\|_F + \eta_\ell\sqrt{n}\,K\sqrt{r}(1+\varepsilon_J)\right]. \tag{S.32}$$

Proof. During round ℓ, the local update and communication step give

$$\mathbf{X}_{\ell+1} = \left(\mathbf{X}_\ell + \sum_{j=0}^{K}\mathbf{e}_{\ell,j}^x - \eta_\ell\sum_{j=0}^{K-1}\mathbf{O}_{\ell,j}\right)\mathbf{W}^\ell.$$

Repeatedly substituting this update and using $\mathbf{X}_0(\mathbf{I}-\mathbf{J})=\mathbf{0}$ gives the following continuous chain:

$$\begin{aligned}
&\mathbb{E}\,\|\mathbf{X}_t(\mathbf{I}-\mathbf{J})\|_F \\
&= \mathbb{E}\left\|\sum_{\ell=0}^{t-1}\left(\sum_{j=0}^{K}\mathbf{e}_{\ell,j}^x - \eta_\ell\sum_{j=0}^{K-1}\mathbf{O}_{\ell,j}\right)(\mathbf{I}-\mathbf{J})\prod_{s=\ell}^{t-1}\mathbf{W}^s\right\|_F \\
&\le \sum_{\ell=0}^{t-1}\mathbb{E}\left\|\left(\sum_{j=0}^{K}\mathbf{e}_{\ell,j}^x - \eta_\ell\sum_{j=0}^{K-1}\mathbf{O}_{\ell,j}\right)(\mathbf{I}-\mathbf{J})\prod_{s=\ell}^{t-1}\mathbf{W}^s\right\|_F \\
&\le \sum_{\ell=0}^{t-1}\mathbb{E}\left\|\sum_{j=0}^{K}\mathbf{e}_{\ell,j}^x - \eta_\ell\sum_{j=0}^{K-1}\mathbf{O}_{\ell,j}\right\|_F\left\|(\mathbf{I}-\mathbf{J})\prod_{s=\ell}^{t-1}\mathbf{W}^s\right\|_{\mathrm{op}} \\
&\le \sum_{\ell=0}^{t-1}\rho^{(t-\ell)/2}\mathbb{E}\left\|\sum_{j=0}^{K}\mathbf{e}_{\ell,j}^x - \eta_\ell\sum_{j=0}^{K-1}\mathbf{O}_{\ell,j}\right\|_F \\
&\le \sum_{\ell=0}^{t-1}\rho^{(t-\ell)/2}\left[\sum_{j=0}^{K}\mathbb{E}\left\|\mathbf{e}_{\ell,j}^x\right\|_F + \eta_\ell\sum_{j=0}^{K-1}\mathbb{E}\,\|\mathbf{O}_{\ell,j}\|_F\right] \\
&= \sum_{\ell=0}^{t-1}\rho^{(t-\ell)/2}\left[\sum_{j=0}^{K}\mathbb{E}\left(\sum_{i=1}^{n}\left\|\mathbf{e}_{i,\ell,j}^x\right\|_F^2\right)^{1/2} + \eta_\ell\sum_{j=0}^{K-1}\mathbb{E}\left(\sum_{i=1}^{n}\|\mathbf{O}_{i,\ell,j}\|_F^2\right)^{1/2}\right] \\
&\le \sum_{\ell=0}^{t-1}\rho^{(t-\ell)/2}\left[\sum_{j=0}^{K}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{e}_{i,\ell,j}^x\right\|_F + \eta_\ell\sqrt{n}\,K\sqrt{r}(1+\varepsilon_J)\right] \\
&\le \sum_{\ell=0}^{t-1}\rho^{(t-\ell)/2}\left[\sum_{j=0}^{K}\sum_{i=1}^{n}\sqrt{q_{i,\ell}^x}\,\mathbb{E}\,\|\mathbf{x}_{i,\ell,j}\|_F + \eta_\ell\sqrt{n}\,K\sqrt{r}(1+\varepsilon_J)\right].
\end{aligned}$$

The first equality unrolls the round recursion, uses $\mathbf{X}_0(\mathbf{I}-\mathbf{J})=\mathbf{0}$, and applies the algebraic identity $(\mathbf{I}-\mathbf{J})\prod_{s=\ell}^{t-1}\mathbf{W}^s = \prod_{s=\ell}^{t-1}(\mathbf{W}^s-\mathbf{J}) = \prod_{s=\ell}^{t-1}\mathbf{W}^s - \mathbf{J}$, which follows from doubly stochastic commutativity $\mathbf{J}\mathbf{W}^s = \mathbf{W}^s\mathbf{J} = \mathbf{J}$ and idempotent projection $\mathbf{J}^2=\mathbf{J}$. The next inequality uses the Frobenius triangle inequality $\|\sum_s A_s\|_F \le \sum_s \|A_s\|_F$. The following inequality uses the mixed-norm bound $\|AB\|_F \le \|A\|_F\|B\|_{\mathrm{op}}$ and Lemma 1. For the next equality, the block-Frobenius norm is expanded as $\|A\|_F^2 = \sum_i \|A_i\|_F^2$. The following inequality uses $(\sum_i a_i^2)^{1/2} \le \sum_i a_i$ and the nonnegative block norms. Lemma 2 supplies the first-moment bound $\mathbb{E}\|\mathbf{e}_{i,\ell,j}^x\|_F \le \sqrt{q_{i,\ell}^x}\,\mathbb{E}\|\mathbf{x}_{i,\ell,j}\|_F$ (by Jensen). Finally, Lemma S.2, whose local bound $\|\mathbf{O}_{i,\ell,j}\|_F \le \sqrt{r}(1+\varepsilon_J)$ follows from Lemma 3 of the main text, gives the update-direction term $\eta_\ell\sqrt{n}K\sqrt{r}(1+\varepsilon_J)$. □

Lemma S.4 (Matrix Hölder Inequality and Polar Duality)**.** *For any matrices* $\mathbf{A},\mathbf{B}\in\mathbb{R}^{m\times d_{\mathrm{m}}}$*:* Matrix Hölder inequality.

$$|\,\langle\mathbf{A},\mathbf{B}\rangle_F\,| \le \|\mathbf{A}\|_*\,\|\mathbf{B}\|_{\mathrm{op}}, \qquad \textit{in particular,} \qquad \langle\mathbf{A},\mathbf{B}\rangle_F \ge -\|\mathbf{A}\|_*\,\|\mathbf{B}\|_{\mathrm{op}}. \tag{S.33}$$

Polar duality. *For any* $\mathbf{X} \in \mathbb{R}^{m\times d_{\mathrm{m}}}$ *with polar factor* $\mathbf{P} \triangleq \mathrm{Polar}(\mathbf{X})$,

$$\langle \mathbf{X}, \mathbf{P}\rangle_F = \|\mathbf{X}\|_*, \qquad \|\mathbf{P}\|_{\mathrm{op}} \le 1. \tag{S.34}$$

Lemma S.5 (Bound on the Network Consensus Drift)**.** *Under Assumptions 1–4, Lemmas 2 and 3 of the main text, and Lemmas S.2 and S.3, with positive non-increasing stepsizes* $\{\eta_t\}_{t=0}^{T-1}$*, nonnegative doubly stochastic mixing matrices, and the initial consensus condition* $\mathbf{X}_0(\mathbf{I}-\mathbf{J}) = \mathbf{0}$*, the cumulative linear network consensus drift satisfies*

$$\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)-\nabla f_i(\mathbf{x}_{i,t,k})\|_F \\
&\le K\sigma_2 + \frac{LK^2\sqrt{r}(1+\varepsilon_J)}{1-\sqrt{\rho}}\frac{\sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t} + \frac{LK}{\sqrt{n}(1-\sqrt{\rho})\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^x}\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F.
\end{aligned} \tag{S.35}$$

Proof. For any round t and step $k \in \{0,\dots,K-1\}$, telescoping the local updates within round t gives $\mathbf{x}_{i,t,k} = \mathbf{x}_{i,t,0} + \sum_{j=0}^{k-1}(\mathbf{e}_{i,t,j}^x - \eta_t\mathbf{O}_{i,t,j})$. Adding and subtracting terms inside the norm, using the triangle inequality, the norm relations $\|M\|_F \le \|M\|_*$ and $\|M\|_{\mathrm{op}} \le \|M\|_F$, and Assumption 1 ($L$-smoothness: $\|\nabla f_i(A)-\nabla f_i(B)\|_F \le \|\nabla f_i(A)-\nabla f_i(B)\|_* \le L\|A-B\|_{\mathrm{op}} \le L\|A-B\|_F$) gives the following continuous master chain:

$$\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)-\nabla f_i(\mathbf{x}_{i,t,k})\|_F \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\left[\nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)-\nabla f_i\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)\right]+\left[\nabla f_i\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)-\nabla f_i(\mathbf{x}_{i,t,0})\right]+\left[\nabla f_i(\mathbf{x}_{i,t,0})-\nabla f_i(\mathbf{x}_{i,t,k})\right]\|_F \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\left[\mathbb{E}\|\nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)-\nabla f_i\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)\|_F + L\mathbb{E}\|\mathbf{X}_t\frac{\mathbf{1}}{n}-\mathbf{x}_{i,t,0}\|_F + L\mathbb{E}\|\mathbf{x}_{i,t,0}-\mathbf{x}_{i,t,k}\|_F\right] \\
&\le K\sigma_2 + \frac{LK}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{X}_t\frac{\mathbf{1}}{n}-\mathbf{x}_{i,t,0}\|_F\right] + \frac{L}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\sum_{j=0}^{k-1}\left(\mathbf{e}_{i,t,j}^x-\eta_t\mathbf{O}_{i,t,j}\right)\|_F \\
&\le K\sigma_2 + \frac{LK}{\sqrt{n}\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\|\mathbf{X}_t(\mathbf{I}-\mathbf{J})\|_F \\
&\quad + \frac{L}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\left[\sum_{j=0}^{k-1}\mathbb{E}\|\mathbf{e}_{i,t,j}^x\|_F + \eta_t\sum_{j=0}^{k-1}\mathbb{E}\|\mathbf{O}_{i,t,j}\|_F\right] \\
&\le K\sigma_2 + \frac{LK}{\sqrt{n}\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{\ell=0}^{t-1}\rho^{(t-\ell)/2}\left[\sum_{j=0}^{K}\sum_{i=1}^{n}\sqrt{q_{i,\ell}^x}\mathbb{E}\|\mathbf{x}_{i,\ell,j}\|_F + \eta_\ell\sqrt{n}K\sqrt{r}(1+\varepsilon_J)\right] \\
&\quad + \frac{L}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\left[K\sqrt{q_{i,t}^x}\sum_{j=0}^{K}\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F + \eta_t\frac{K^2}{2}\sqrt{r}(1+\varepsilon_J)\right] \\
&\le K\sigma_2 + \frac{LK^2\sqrt{r}(1+\varepsilon_J)}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{\ell=0}^{t-1}\rho^{(t-\ell)/2}\eta_\ell + \frac{LK^2\sqrt{r}(1+\varepsilon_J)}{2\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t^2 \\
&\quad + \frac{LK}{\sqrt{n}\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{\ell=0}^{t-1}\rho^{(t-\ell)/2}\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,\ell}^x}\mathbb{E}\|\mathbf{x}_{i,\ell,j}\|_F \\
&\quad + \frac{LK}{n\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^x}\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F \\
&\le K\sigma_2 + LK^2\sqrt{r}(1+\varepsilon_J)\left(\frac{\sqrt{\rho}}{1-\sqrt{\rho}}+\frac{1}{2}\right)\frac{\sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t} \\
&\quad + \frac{LK\sqrt{\rho}}{\sqrt{n}(1-\sqrt{\rho})\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^x}\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{LK}{n\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^{x}}\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F \\
&< K\sigma_2 + \frac{LK^2\sqrt{r}(1+\varepsilon_J)}{1-\sqrt{\rho}}\frac{\sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t} \\
&\quad + LK\left(\frac{\sqrt{\rho}}{\sqrt{n}(1-\sqrt{\rho})}+\frac{1}{n}\right)\frac{\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^{x}}\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F}{\sum_{t=0}^{T-1}\eta_t} \\
&\le K\sigma_2 + \frac{LK^2\sqrt{r}(1+\varepsilon_J)}{1-\sqrt{\rho}}\frac{\sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t} + \frac{LK}{\sqrt{n}(1-\sqrt{\rho})\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^{x}}\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F. \qquad \text{(S.36)}
\end{aligned}
$$

In Eq. (S.36), the first equality is the exact decomposition obtained by adding and subtracting $\nabla f_i(\mathbf{X}_t\frac{\mathbf{1}}{n})$ and $\nabla f_i(\mathbf{x}_{i,t,0})$. The next inequality uses the triangle inequality, the norm relations $\|M\|_F \le \|M\|_*$ and $\|M\|_{\mathrm{op}} \le \|M\|_F$, and Assumption 1 ($L$-smoothness: $\|\nabla f_i(A) - \nabla f_i(B)\|_F \le \|\nabla f_i(A) - \nabla f_i(B)\|_* \le L\|A-B\|_{\mathrm{op}} \le L\|A-B\|_F$). The following inequality bounds the client heterogeneity term via Assumption 3 ($\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f(\mathbf{X}_t\frac{\mathbf{1}}{n}) - \nabla f_i(\mathbf{X}_t\frac{\mathbf{1}}{n})\|_F \le [\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\|\nabla f(\mathbf{X}_t\frac{\mathbf{1}}{n}) - \nabla f_i(\mathbf{X}_t\frac{\mathbf{1}}{n})\|_F)^2]^{1/2} \le \sigma_2$), and applies the Cauchy–Schwarz inequality over clients $\frac{1}{n}\sum_{i=1}^{n}a_i \le \frac{1}{\sqrt{n}}(\sum_{i=1}^{n}a_i^2)^{1/2}$ together with the block-Frobenius identity $\|\mathbf{X}_t(\mathbf{I}-\mathbf{J})\|_F^2 = \sum_{i=1}^{n}\|\mathbf{X}_t\frac{\mathbf{1}}{n} - \mathbf{x}_{i,t,0}\|_F^2$ to obtain $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{X}_t\frac{\mathbf{1}}{n} - \mathbf{x}_{i,t,0}\|_F \le \frac{1}{\sqrt{n}}\mathbb{E}\|\mathbf{X}_t(\mathbf{I}-\mathbf{J})\|_F$. The intra-round consensus term is bounded via Lemma S.3. The local drift sum is exchanged as $\sum_{k=0}^{K-1}\sum_{j=0}^{k-1}a_j = \sum_{j=0}^{K-2}(K-1-j)a_j$. The double sum over rounds is bounded using the non-increasing stepsize property $\eta_t \le \eta_\ell$ for $\ell \le t$: $\sum_{t=0}^{T-1}\eta_t\sum_{\ell=0}^{t-1}\rho^{(t-\ell)/2}\eta_\ell = \sum_{\ell=0}^{T-2}\eta_\ell\sum_{t=\ell+1}^{T-1}\rho^{(t-\ell)/2}\eta_t \le \sum_{\ell=0}^{T-2}\eta_\ell^2\sum_{s=1}^{\infty}\rho^{s/2} = \frac{\sqrt{\rho}}{1-\sqrt{\rho}}\sum_{t=0}^{T-1}\eta_t^2$. Combining $\frac{1}{2} + \frac{\sqrt{\rho}}{1-\sqrt{\rho}} < \frac{1}{1-\sqrt{\rho}}$ gives the stepsize coefficient. Finally, bounding $K-1-j \le K$ and $\frac{\sqrt{\rho}}{\sqrt{n}(1-\sqrt{\rho})} + \frac{1}{n} \le \frac{\sqrt{\rho}+1-\sqrt{\rho}}{\sqrt{n}(1-\sqrt{\rho})} = \frac{1}{\sqrt{n}(1-\sqrt{\rho})}$ yields Eq. (S.35). □

Lemma S.6 (Bound on the Stochastic Noise Tracking Error)**.** *Under Assumption 2, with $0 \le \beta < 1$ and positive non-increasing stepsizes $\{\eta_t\}_{t=0}^{T-1}$, and the boundary conditions $\mathbf{C}_{i,0,0} = \mathbf{S}_{i,0,0} = \mathbf{0}$ and $\mathbf{C}_{i,t,0} = \mathbf{C}_{i,t-1,K}$, $\mathbf{S}_{i,t,0} = \mathbf{S}_{i,t-1,K}$ for $t \ge 1$, the cumulative stochastic noise tracking error for momentum buffers $\mathbf{C}_{i,t,k}$ and $\mathbf{S}_{i,t,k}$ across global rounds $t \in \{0,\dots,T-1\}$ and local steps $k \in \{0,\dots,K-1\}$ satisfies:*

$$
\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{C}_{i,t,k+1} - \mathbf{S}_{i,t,k+1}\|_F \le K\sigma_1\sqrt{\frac{1-\beta}{1+\beta}}. \qquad \text{(S.37)}
$$

Proof.

$$
\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{C}_{i,t,k+1} - \mathbf{S}_{i,t,k+1}\|_F \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\sqrt{\mathbb{E}\|\mathbf{S}_{i,t,k+1} - \mathbf{C}_{i,t,k+1}\|_F^2} \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\sqrt{\mathbb{E}\|\beta(\mathbf{S}_{i,t,k} - \mathbf{C}_{i,t,k}) + (1-\beta)\big(\nabla F_i(\mathbf{x}_{i,t,k};\xi_{i,t,k}) - \nabla f_i(\mathbf{x}_{i,t,k})\big)\|_F^2} \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Bigg(\beta^2\mathbb{E}\|\mathbf{S}_{i,t,k} - \mathbf{C}_{i,t,k}\|_F^2 + (1-\beta)^2\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,t,k};\xi_{i,t,k}) - \nabla f_i(\mathbf{x}_{i,t,k})\|_F^2 \\
&\qquad + 2\beta(1-\beta)\mathbb{E}\left\langle \mathbf{S}_{i,t,k} - \mathbf{C}_{i,t,k}, \nabla F_i(\mathbf{x}_{i,t,k};\xi_{i,t,k}) - \nabla f_i(\mathbf{x}_{i,t,k})\right\rangle_F\Bigg)^{1/2} \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\sqrt{\beta^2\mathbb{E}\|\mathbf{S}_{i,t,k} - \mathbf{C}_{i,t,k}\|_F^2 + (1-\beta)^2\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,t,k};\xi_{i,t,k}) - \nabla f_i(\mathbf{x}_{i,t,k})\|_F^2}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Bigg(\beta^{2(k+1)}\mathbb{E}\|\mathbf{S}_{i,t,0}-\mathbf{C}_{i,t,0}\|_F^2 \\
&\quad+\sum_{j=0}^{k}(1-\beta)^2\beta^{2(k-j)}\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,t,j};\xi_{i,t,j})-\nabla f_i(\mathbf{x}_{i,t,j})\|_F^2\Bigg)^{1/2} \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Bigg(\beta^{2(k+1+tK)}\mathbb{E}\|\mathbf{S}_{i,0,0}-\mathbf{C}_{i,0,0}\|_F^2 \\
&\quad+\sum_{\tau=0}^{t-1}\beta^{2(k+1+(t-1-\tau)K)}\sum_{j=0}^{K-1}(1-\beta)^2\beta^{2(K-1-j)}\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,\tau,j};\xi_{i,\tau,j})-\nabla f_i(\mathbf{x}_{i,\tau,j})\|_F^2 \\
&\quad+\sum_{j=0}^{k}(1-\beta)^2\beta^{2(k-j)}\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,t,j};\xi_{i,t,j})-\nabla f_i(\mathbf{x}_{i,t,j})\|_F^2\Bigg)^{1/2} \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Bigg(\sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)^2\beta^{2(k+(t-\tau)K-j)}\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,\tau,j};\xi_{i,\tau,j})-\nabla f_i(\mathbf{x}_{i,\tau,j})\|_F^2 \\
&\quad+\sum_{j=0}^{k}(1-\beta)^2\beta^{2(k-j)}\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,t,j};\xi_{i,t,j})-\nabla f_i(\mathbf{x}_{i,t,j})\|_F^2\Bigg)^{1/2} \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\left(\sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)^2\beta^{2(k+(t-\tau)K-j)}\sigma_1^2+\sum_{j=0}^{k}(1-\beta)^2\beta^{2(k-j)}\sigma_1^2\right)^{1/2} \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\sqrt{(1-\beta)^2\sigma_1^2\sum_{l=0}^{tK+k}\beta^{2l}} \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\sqrt{(1-\beta)^2\sigma_1^2\sum_{l=0}^{\infty}\beta^{2l}} \\
&= \frac{K\sigma_1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sqrt{\frac{1-\beta}{1+\beta}} \\
&= K\sigma_1\sqrt{\frac{1-\beta}{1+\beta}}. \qquad \text{(S.38)}
\end{aligned}
$$

The recursion in the third line follows by subtracting (S.25) from (S.26): $\mathbf{S}_{i,t,k+1}-\mathbf{C}_{i,t,k+1}=\beta(\mathbf{S}_{i,t,k}-\mathbf{C}_{i,t,k})+(1-\beta)[\nabla F_i(\mathbf{x}_{i,t,k};\xi_{i,t,k})-\nabla f_i(\mathbf{x}_{i,t,k})]$. By Assumption 2, $\mathbb{E}[\|\nabla F_i(\mathbf{x}_{i,\tau,j};\xi_{i,\tau,j})-\nabla f_i(\mathbf{x}_{i,\tau,j})\|_F^2]\le\sigma_1^2$, and the cross term $2\beta(1-\beta)\mathbb{E}\langle\mathbf{S}_{i,t,k}-\mathbf{C}_{i,t,k},\nabla F_i(\mathbf{x}_{i,t,k};\xi_{i,t,k})-\nabla f_i(\mathbf{x}_{i,t,k})\rangle_F$ vanishes because $\mathbf{S}_{i,t,k}-\mathbf{C}_{i,t,k}$ is measurable with respect to the history $\mathcal{H}$ before the k-th oracle call ($\mathcal{H}$ is the σ-algebra of all randomness realized before that call) and $\mathbb{E}[\nabla F_i(\mathbf{x}_{i,t,k};\xi_{i,t,k})-\nabla f_i(\mathbf{x}_{i,t,k})\mid\mathcal{H}]=\mathbf{0}$ by the conditional unbiasedness in Assumption 2. Since $0\le\beta<1$, $\sum_{l=0}^{\infty}\beta^{2l}=1/(1-\beta^2)$, and $(1-\beta)\sigma_1/\sqrt{1-\beta^2}=\sigma_1\sqrt{(1-\beta)/(1+\beta)}$. Averaging over rounds and clients completes the proof of Eq. (S.37). □

Lemma S.7 (Bound on the True Gradient Tracking Bias)**.** *Under Assumptions 1–4, Lemmas 2 and 3 of the main text, and Lemmas S.1, S.2, and S.5, with $0\le\beta<1$, positive non-increasing stepsizes $\{\eta_t\}_{t=0}^{T-1}$, and $\mathbf{C}_{i,0,0}=\mathbf{0}$ for every i, the cumulative true gradient tracking bias obeys*

$$
\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,k})-\mathbf{C}_{i,t,k+1}\|_F \\
&\le 2\beta K\sigma_2+\frac{\beta}{(1-\beta)\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\,\mathbb{E}\left\|\nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)\right\|_*
\end{aligned}
$$

$$+ L\beta K\sqrt{r}(1+\varepsilon_J)\left[\frac{2K}{1-\sqrt{\rho}}+\frac{1}{1-\beta}\right]\frac{\sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t}$$
$$+ \frac{L\beta}{\sum_{t=0}^{T-1}\eta_t}\left[\frac{2K}{\sqrt{n}(1-\sqrt{\rho})}+\frac{1}{n(1-\beta)}\right]\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^x}\,\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F. \tag{S.39}$$

Proof. Unrolling the momentum buffer $\mathbf{C}_{i,t,k+1} = \beta\mathbf{C}_{i,t,k} + (1-\beta)\nabla f_i(\mathbf{x}_{i,t,k})$ with boundary condition $\mathbf{C}_{i,t,0} = \mathbf{C}_{i,t-1,K}$ and initialization $\mathbf{C}_{i,0,0} = \mathbf{0}$, and using the geometric partition $(1-\beta)\sum_{l=0}^{tK+k}\beta^l + \beta^{tK+k+1} = 1$, gives the following continuous chain. The average-iterate differences are expanded directly in the chain using Lemma S.1, Lemma 2, and Lemma S.2.

$$\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,k}) - \mathbf{C}_{i,t,k+1}\|_F \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_{i,t,k}) - \left(\beta^{k+1}\mathbf{C}_{i,t,0} + \sum_{j=0}^{k}(1-\beta)\beta^{k-j}\nabla f_i(\mathbf{x}_{i,t,j})\right)\right\|_F \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\Bigg\|\nabla f_i(\mathbf{x}_{i,t,k}) - \Bigg(\beta^{k+1+tK}\mathbf{C}_{i,0,0} \\
&\quad + \sum_{\tau=0}^{t-1}\beta^{k+1+(t-1-\tau)K}\sum_{j=0}^{K-1}(1-\beta)\beta^{K-1-j}\nabla f_i(\mathbf{x}_{i,\tau,j}) + \sum_{j=0}^{k}(1-\beta)\beta^{k-j}\nabla f_i(\mathbf{x}_{i,t,j})\Bigg)\Bigg\|_F \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\Bigg\|\beta^{tK+k+1}\left(\nabla f_i(\mathbf{x}_{i,t,k}) - \mathbf{C}_{i,0,0}\right) \\
&\quad + (1-\beta)\Bigg[\sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}\beta^{k+(t-\tau)K-j}\left(\nabla f_i(\mathbf{x}_{i,t,k}) - \nabla f_i(\mathbf{x}_{i,\tau,j})\right) \\
&\qquad + \sum_{j=0}^{k-1}\beta^{k-j}\left(\nabla f_i(\mathbf{x}_{i,t,k}) - \nabla f_i(\mathbf{x}_{i,t,j})\right)\Bigg]\Bigg\|_F \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\Bigg\|\sum_{j=0}^{k-1}(1-\beta)\beta^{k-j}\big(\nabla f_i(\mathbf{x}_{i,t,k}) - \nabla f_i(\mathbf{x}_{i,t,j})\big) \\
&\quad + \sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\big(\nabla f_i(\mathbf{x}_{i,t,k}) - \nabla f_i(\mathbf{x}_{i,\tau,j})\big) + \beta^{tK+k+1}\nabla f_i(\mathbf{x}_{i,t,k})\Bigg\|_F \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\Bigg\|\sum_{j=0}^{k-1}(1-\beta)\beta^{k-j}\Big([\nabla f_i(\mathbf{x}_{i,t,k}) - \nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})] + [\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n}) - \nabla f_i(\mathbf{x}_{i,t,j})]\Big) \\
&\quad + \sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\Big([\nabla f_i(\mathbf{x}_{i,t,k}) - \nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})] + [\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n}) - \nabla f(\mathbf{X}_\tau\tfrac{\mathbf{1}}{n})] \\
&\qquad + [\nabla f(\mathbf{X}_\tau\tfrac{\mathbf{1}}{n}) - \nabla f_i(\mathbf{x}_{i,\tau,j})]\Big) + \beta^{tK+k+1}\Big(\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n}) + [\nabla f_i(\mathbf{x}_{i,t,k}) - \nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})]\Big)\Bigg\|_F \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Bigg[\sum_{j=0}^{k-1}(1-\beta)\beta^{k-j}\Big(\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,k}) - \nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\|_F + \mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,j}) - \nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\|_F\Big) \\
&\quad + \sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\Big(\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,k}) - \nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\|_F + L\mathbb{E}\|\mathbf{X}_t\tfrac{\mathbf{1}}{n} - \mathbf{X}_\tau\tfrac{\mathbf{1}}{n}\|_F \\
&\qquad + \mathbb{E}\|\nabla f(\mathbf{X}_\tau\tfrac{\mathbf{1}}{n}) - \nabla f_i(\mathbf{x}_{i,\tau,j})\|_F\Big) + \beta^{tK+k+1}\Big(\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_* + \mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,k}) - \nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\|_F\Big)\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Bigg[\left(\sum_{j=0}^{k-1}(1-\beta)\beta^{k-j}+\sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}+\beta^{tK+k+1}\right)\\
&\qquad\times\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,k})-\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\|_F+\sum_{j=0}^{k-1}(1-\beta)\beta^{k-j}\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,j})-\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\|_F\\
&\quad+\sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\mathbb{E}\|\nabla f(\mathbf{X}_\tau\tfrac{\mathbf{1}}{n})-\nabla f_i(\mathbf{x}_{i,\tau,j})\|_F+\beta^{tK+k+1}\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_*\\
&\quad+L\sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\mathbb{E}\|\mathbf{X}_t\tfrac{\mathbf{1}}{n}-\mathbf{X}_\tau\tfrac{\mathbf{1}}{n}\|_F\Bigg]\\
&=\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Bigg[\left((1-\beta)\sum_{\ell=1}^{k}\beta^{\ell}+(1-\beta)\sum_{\ell=k+1}^{tK+k}\beta^{\ell}+\beta^{tK+k+1}\right)\\
&\qquad\times\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,k})-\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\|_F+\sum_{j=0}^{k-1}(1-\beta)\beta^{k-j}\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,j})-\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\|_F\\
&\quad+\sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\mathbb{E}\|\nabla f(\mathbf{X}_\tau\tfrac{\mathbf{1}}{n})-\nabla f_i(\mathbf{x}_{i,\tau,j})\|_F+\beta^{tK+k+1}\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_*\\
&\quad+L\sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\mathbb{E}\|\mathbf{X}_t\tfrac{\mathbf{1}}{n}-\mathbf{X}_\tau\tfrac{\mathbf{1}}{n}\|_F\Bigg]\\
&=\frac{\beta}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,k})-\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\|_F\\
&\quad+\frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\sum_{v=0}^{K-1}\mathbb{E}\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})-\nabla f_i(\mathbf{x}_{i,u,v})\|_F\\
&\quad\times\left[\eta_u\sum_{k=v+1}^{K-1}(1-\beta)\beta^{k-v}+\sum_{t=u+1}^{T-1}\eta_t\sum_{k=0}^{K-1}(1-\beta)\beta^{k+(t-u)K-v}\right]\\
&\quad+\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_*\sum_{k=0}^{K-1}\beta^{tK+k+1}\\
&\quad+\frac{L}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{\tau=0}^{t-1}\sum_{k=0}^{K-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\mathbb{E}\|\mathbf{X}_t\tfrac{\mathbf{1}}{n}-\mathbf{X}_\tau\tfrac{\mathbf{1}}{n}\|_F\\
&\leq\frac{\beta}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,k})-\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\|_F\\
&\quad+\frac{\beta}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\mathbb{E}\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})-\nabla f_i(\mathbf{x}_{i,u,v})\|_F\\
&\quad+\frac{\beta}{(1-\beta)\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_*\\
&\quad+\frac{L}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{\tau=0}^{t-1}\sum_{k=0}^{K-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\mathbb{E}\|\mathbf{X}_t\tfrac{\mathbf{1}}{n}-\mathbf{X}_\tau\tfrac{\mathbf{1}}{n}\|_F\\
&\leq 2\beta\left(\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})-\nabla f_i(\mathbf{x}_{i,t,k})\|_F\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{\beta}{(1-\beta)\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_* \\
&+\frac{L}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{\tau=0}^{t-1}\sum_{k=0}^{K-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\sum_{s=\tau}^{t-1}\left[\frac{1}{n}\sum_{a=1}^{n}\sum_{\ell=0}^{K}\sqrt{q_{a,s}^x}\,\mathbb{E}\|\mathbf{x}_{a,s,\ell}\|_F+\eta_sK\sqrt{r}(1+\varepsilon_J)\right] \\
=\;&2\beta\left(\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})-\nabla f_i(\mathbf{x}_{i,t,k})\|_F\right) \\
&+\frac{\beta}{(1-\beta)\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_* \\
&+\frac{L}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{\tau=0}^{t-1}\frac{\beta^{(t-\tau-1)K+1}(1-\beta^K)^2}{1-\beta}\sum_{s=\tau}^{t-1}\left[\frac{1}{n}\sum_{a=1}^{n}\sum_{\ell=0}^{K}\sqrt{q_{a,s}^x}\,\mathbb{E}\|\mathbf{x}_{a,s,\ell}\|_F+\eta_sK\sqrt{r}(1+\varepsilon_J)\right] \\
=\;&2\beta\left(\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})-\nabla f_i(\mathbf{x}_{i,t,k})\|_F\right) \\
&+\frac{\beta}{(1-\beta)\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_* \\
&+\frac{L}{\sum_{t=0}^{T-1}\eta_t}\sum_{s=0}^{T-2}\left[\frac{1}{n}\sum_{a=1}^{n}\sum_{\ell=0}^{K}\sqrt{q_{a,s}^x}\,\mathbb{E}\|\mathbf{x}_{a,s,\ell}\|_F+\eta_sK\sqrt{r}(1+\varepsilon_J)\right] \\
&\quad\times\sum_{t=s+1}^{T-1}\eta_t\sum_{\tau=0}^{s}\frac{\beta^{(t-\tau-1)K+1}(1-\beta^K)^2}{1-\beta} \\
\le\;&2\beta\left(\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})-\nabla f_i(\mathbf{x}_{i,t,k})\|_F\right) \\
&+\frac{\beta}{(1-\beta)\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_* \\
&+\frac{L\beta}{(1-\beta)\sum_{t=0}^{T-1}\eta_t}\sum_{s=0}^{T-1}\eta_s\left[\frac{1}{n}\sum_{a=1}^{n}\sum_{\ell=0}^{K}\sqrt{q_{a,s}^x}\,\mathbb{E}\|\mathbf{x}_{a,s,\ell}\|_F+\eta_sK\sqrt{r}(1+\varepsilon_J)\right] \\
\le\;&2\beta\Bigg[K\sigma_2+\frac{LK^2\sqrt{r}(1+\varepsilon_J)}{1-\sqrt{\rho}}\frac{\sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t} \\
&+\frac{LK}{\sqrt{n}(1-\sqrt{\rho})\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^x}\,\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F\Bigg] \\
&+\frac{\beta}{(1-\beta)\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_* \\
&+\frac{L\beta}{(1-\beta)\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\left[\frac{1}{n}\sum_{a=1}^{n}\sum_{\ell=0}^{K}\sqrt{q_{a,t}^x}\,\mathbb{E}\|\mathbf{x}_{a,t,\ell}\|_F+\eta_tK\sqrt{r}(1+\varepsilon_J)\right] \\
=\;&2\beta K\sigma_2+\frac{\beta}{(1-\beta)\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_* \\
&+L\beta K\sqrt{r}(1+\varepsilon_J)\left[\frac{2K}{1-\sqrt{\rho}}+\frac{1}{1-\beta}\right]\frac{\sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t} \\
&+\frac{L\beta}{\sum_{t=0}^{T-1}\eta_t}\left[\frac{2K}{\sqrt{n}(1-\sqrt{\rho})}+\frac{1}{n(1-\beta)}\right]\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^x}\,\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F. \qquad \text{(S.40)}
\end{aligned}
$$

□

Lemma S.8 (Bound on the Stochastic Gradient Quantization Error)**.** *Under Assumptions 1–4, Lemmas 2 and 3 of the main text, and Lemmas S.3 and S.2, for $0 \le \beta < 1$ and positive non-increasing stepsizes $\{\eta_t\}_{t=0}^{T-1}$, the cumulative stochastic gradient quantization error satisfies*

$$\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{S}_{i,t,k+1}-\mathbf{Y}_{i,t,k+1}\|_F \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sqrt{q_{i,t}^{g}}\Bigg[K\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_* + \sigma_1+\sigma_2\right) \\
&\qquad + LK\sum_{\ell=0}^{t}\rho^{(t-\ell)/2}\left(\sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q_{a,\ell}^{x}}\,\mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K\sqrt{nr}(1+\varepsilon_J)\right)\Bigg]. \qquad \text{(S.41)}
\end{aligned}$$

Proof. Unrolling the two momentum recursions and applying the triangle inequality gives

$$\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{S}_{i,t,k+1}-\mathbf{Y}_{i,t,k+1}\|_F \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\Bigg\|\sum_{j=0}^{k}(1-\beta)\beta^{k-j}\left(\nabla F_i(\mathbf{x}_{i,t,j};\xi_{i,t,j})-\nabla^{Q}F_i(\mathbf{x}_{i,t,j};\xi_{i,t,j})\right) \\
&\qquad + \sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\left(\nabla F_i(\mathbf{x}_{i,\tau,j};\xi_{i,\tau,j})-\nabla^{Q}F_i(\mathbf{x}_{i,\tau,j};\xi_{i,\tau,j})\right)\Bigg\|_F \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Bigg[\sum_{j=0}^{k}(1-\beta)\beta^{k-j}\sqrt{q_{i,t}^{g}}\,\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,t,j};\xi_{i,t,j})\|_F \\
&\qquad + \sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}(1-\beta)\beta^{k+(t-\tau)K-j}\sqrt{q_{i,\tau}^{g}}\,\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,\tau,j};\xi_{i,\tau,j})\|_F\Bigg] \\
&= \tfrac{1}{\sum_{t=0}^{T-1}\eta_t}\tfrac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\sum_{v=0}^{K-1}\sqrt{q_{i,u}^{g}}\,\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,u,v};\xi_{i,u,v})\|_F\left[\eta_u\sum_{k=v}^{K-1}(1-\beta)\beta^{k-v}+\sum_{t=u+1}^{T-1}\eta_t\sum_{k=0}^{K-1}(1-\beta)\beta^{k+(t-u)K-v}\right] \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\sqrt{q_{i,u}^{g}}\,\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,u,v};\xi_{i,u,v})\|_F \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\sqrt{q_{i,u}^{g}}\Bigg[\sigma_1+\mathbb{E}\|\nabla f_i(\mathbf{x}_{i,u,v})-\nabla f_i(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\|_F \\
&\qquad + \mathbb{E}\|\nabla f_i(\mathbf{X}_u\tfrac{\mathbf{1}}{n})-\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\|_F+\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_*\Bigg] \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\sqrt{q_{i,u}^{g}}\left[\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_*+\sigma_1+\sigma_2+L\mathbb{E}\|\mathbf{x}_{i,u,v}-\mathbf{X}_u\tfrac{\mathbf{1}}{n}\|_F\right] \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sqrt{q_{i,u}^{g}}\left[K\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_*+\sigma_1+\sigma_2\right)+L\sum_{v=0}^{K-1}\mathbb{E}\|\mathbf{x}_{i,u,v}-\mathbf{X}_u\tfrac{\mathbf{1}}{n}\|_F\right] \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sqrt{q_{i,u}^{g}}\Bigg[K\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_*+\sigma_1+\sigma_2\right)+L\sum_{v=0}^{K-1}\sum_{j=0}^{v-1}\left(\mathbb{E}\|\mathbf{e}_{i,u,j}^{x}\|_F+\eta_u\mathbb{E}\|\mathbf{O}_{i,u,j}\|_F\right) \\
&\qquad + LK\mathbb{E}\|\mathbf{X}_u(\mathbf{I}-\mathbf{J})\|_F\Bigg] \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sqrt{q_{i,u}^{g}}\Bigg[K\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_*+\sigma_1+\sigma_2\right)
\end{aligned}$$

$$
\begin{aligned}
&\quad + L\sqrt{q_{i,u}^x}\sum_{j=0}^{K-2}(K-1-j)\mathbb{E}\|\mathbf{x}_{i,u,j}\|_F + \frac{LK(K-1)}{2}\eta_u\sqrt{r}(1+\varepsilon_J) \\
&\quad + LK\sum_{\ell=0}^{u-1}\rho^{(u-\ell)/2}\left[\sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q_{a,\ell}^x}\,\mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K\sqrt{nr}(1+\varepsilon_J)\right]\Bigg] \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sqrt{q_{i,u}^g}\Bigg[K\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_* + \sigma_1 + \sigma_2\right) \\
&\quad + LK\left[\sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q_{a,u}^x}\,\mathbb{E}\|\mathbf{x}_{a,u,j}\|_F + \eta_u K\sqrt{nr}(1+\varepsilon_J)\right] \\
&\quad + LK\sum_{\ell=0}^{u-1}\rho^{(u-\ell)/2}\left[\sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q_{a,\ell}^x}\,\mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K\sqrt{nr}(1+\varepsilon_J)\right]\Bigg] \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sqrt{q_{i,u}^g}\Bigg[K\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_* + \sigma_1 + \sigma_2\right) \\
&\quad + LK\sum_{\ell=0}^{u}\rho^{(u-\ell)/2}\left[\sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q_{a,\ell}^x}\,\mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K\sqrt{nr}(1+\varepsilon_J)\right]\Bigg]. \qquad \text{(S.42)}
\end{aligned}
$$

□

Lemma S.9 (Bound on the Forward Weight Quantization Error)**.** *Under Assumptions 1–4, Lemmas 2 and 3 of the main text, and Lemmas S.3 and S.2, with positive non-increasing stepsizes* $\{\eta_t\}_{t=0}^{T-1}$, $0 \le \beta < 1$, *the initial consensus condition* $\mathbf{X}_0(\mathbf{I}-\mathbf{J}) = \mathbf{0}$, *and* $\mathbf{Y}_{i,0,0} = \mathbf{Z}_{i,0,0} = \mathbf{0}$ *for every* i, *the cumulative forward weight quantization error satisfies*

$$
\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{Y}_{i,t,k+1} - \mathbf{Z}_{i,t,k+1}\|_F \\
&\le \frac{K}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\Bigg[2\sqrt{q_{i,t}^g}\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_* + \sigma_1 + \sigma_2\right) + 2\sigma_1 \\
&\quad + L\left(2\sqrt{q_{i,t}^g} + \frac{1+\sqrt{q_{i,t}^g}}{K}\right)\sum_{\ell=0}^{t}\rho^{(t-\ell)/2}\left[\sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q_{a,\ell}^x}\,\mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K\sqrt{nr}(1+\varepsilon_J)\right]\Bigg]. \qquad \text{(S.43)}
\end{aligned}
$$

The convolution is written as a single merged sum over $\ell = 0, \dots, t$, *so the current-round term* $\ell = t$ *is included.*

Proof. By recursive expansion of the updates $\mathbf{Y}_{i,t,k+1}$ and $\mathbf{Z}_{i,t,k+1}$ and the boundary condition $\mathbf{Y}_{i,0,0} = \mathbf{Z}_{i,0,0} = \mathbf{0}$, we obtain the following continuous chain:

$$
\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{Y}_{i,t,k+1} - \mathbf{Z}_{i,t,k+1}\|_F \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\Bigg\|(1-\beta)\sum_{j=0}^{k}\beta^{k-j}\left(\nabla^Q F_i(\mathbf{x}_{i,t,j};\xi_{i,t,j}) - \nabla^Q F_i(\mathbf{x}_{i,t,j}^Q;\xi_{i,t,j})\right) \\
&\qquad + (1-\beta)\sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}\beta^{k+1+(t-1-\tau)K+K-1-j}\left(\nabla^Q F_i(\mathbf{x}_{i,\tau,j};\xi_{i,\tau,j}) - \nabla^Q F_i(\mathbf{x}_{i,\tau,j}^Q;\xi_{i,\tau,j})\right)\Bigg\|_F \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Bigg[(1-\beta)\sum_{j=0}^{k}\beta^{k-j}\mathbb{E}\|\nabla^Q F_i(\mathbf{x}_{i,t,j};\xi_{i,t,j}) - \nabla^Q F_i(\mathbf{x}_{i,t,j}^Q;\xi_{i,t,j})\|_F \\
&\qquad + (1-\beta)\sum_{\tau=0}^{t-1}\sum_{j=0}^{K-1}\beta^{k+1+(t-1-\tau)K+K-1-j}\mathbb{E}\|\nabla^Q F_i(\mathbf{x}_{i,\tau,j};\xi_{i,\tau,j}) - \nabla^Q F_i(\mathbf{x}_{i,\tau,j}^Q;\xi_{i,\tau,j})\|_F\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\sum_{v=0}^{K-1}\mathbb{E}\|\nabla^Q F_i(\mathbf{x}_{i,u,v};\xi_{i,u,v}) - \nabla^Q F_i(\mathbf{x}^Q_{i,u,v};\xi_{i,u,v})\|_F \\
&\qquad \times \left[\eta_u(1-\beta)\sum_{k=v}^{K-1}\beta^{k-v} + \sum_{t=u+1}^{T-1}\eta_t(1-\beta)\sum_{k=0}^{K-1}\beta^{k+1+(t-1-u)K+K-1-v}\right] \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\mathbb{E}\|\nabla^Q F_i(\mathbf{x}_{i,u,v};\xi_{i,u,v}) - \nabla^Q F_i(\mathbf{x}^Q_{i,u,v};\xi_{i,u,v})\|_F \\
&= \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\mathbb{E}\Big\|\Big[\nabla^Q F_i(\mathbf{x}_{i,u,v};\xi_{i,u,v}) - \nabla F_i(\mathbf{x}_{i,u,v};\xi_{i,u,v})\Big] \\
&\qquad + \Big[\nabla F_i(\mathbf{x}_{i,u,v};\xi_{i,u,v}) - \nabla f_i(\mathbf{x}_{i,u,v})\Big] + \Big[\nabla f_i(\mathbf{x}_{i,u,v}) - \nabla f_i(\mathbf{x}^Q_{i,u,v})\Big] \\
&\qquad + \Big[\nabla f_i(\mathbf{x}^Q_{i,u,v}) - \nabla F_i(\mathbf{x}^Q_{i,u,v};\xi_{i,u,v})\Big] + \Big[\nabla F_i(\mathbf{x}^Q_{i,u,v};\xi_{i,u,v}) - \nabla^Q F_i(\mathbf{x}^Q_{i,u,v};\xi_{i,u,v})\Big]\Big\|_F \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\Bigg[\mathbb{E}\|\nabla^Q F_i(\mathbf{x}_{i,u,v};\xi_{i,u,v}) - \nabla F_i(\mathbf{x}_{i,u,v};\xi_{i,u,v})\|_F \\
&\qquad + \mathbb{E}\|\nabla F_i(\mathbf{x}_{i,u,v};\xi_{i,u,v}) - \nabla f_i(\mathbf{x}_{i,u,v})\|_F + L\mathbb{E}\|\mathbf{x}_{i,u,v} - \mathbf{x}^Q_{i,u,v}\|_F \\
&\qquad + \mathbb{E}\|\nabla f_i(\mathbf{x}^Q_{i,u,v}) - \nabla F_i(\mathbf{x}^Q_{i,u,v};\xi_{i,u,v})\|_F + \mathbb{E}\|\nabla F_i(\mathbf{x}^Q_{i,u,v};\xi_{i,u,v}) - \nabla^Q F_i(\mathbf{x}^Q_{i,u,v};\xi_{i,u,v})\|_F\Bigg] \\
&\le \frac{1}{\sum_{t=0}^{T-1}\eta_t}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\Bigg[\sqrt{q^g_{i,u}}\,\mathbb{E}\|\nabla F_i(\mathbf{x}_{i,u,v};\xi_{i,u,v})\|_F + \sigma_1 + L\sqrt{q^x_{i,u}}\,\mathbb{E}\|\mathbf{x}_{i,u,v}\|_F \\
&\qquad + \sigma_1 + \sqrt{q^g_{i,u}}\,\mathbb{E}\|\nabla F_i(\mathbf{x}^Q_{i,u,v};\xi_{i,u,v})\|_F\Bigg] \\
&\le \frac{1}{\sum_{u=0}^{T-1}\eta_u}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\Bigg[2\sqrt{q^g_{i,u}}\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_* + \sigma_1 + \sigma_2\right) + 2\sigma_1 \\
&\qquad + 2\sqrt{q^g_{i,u}}L\mathbb{E}\|\mathbf{x}_{i,u,v} - \mathbf{X}_u\tfrac{\mathbf{1}}{n}\|_F + (1+\sqrt{q^g_{i,u}})L\sqrt{q^x_{i,u}}\mathbb{E}\|\mathbf{x}_{i,u,v}\|_F\Bigg] \\
&= \frac{1}{\sum_{u=0}^{T-1}\eta_u}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\Bigg[2\sqrt{q^g_{i,u}}\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_* + \sigma_1 + \sigma_2\right) + 2\sigma_1 \\
&\qquad + 2\sqrt{q^g_{i,u}}L\mathbb{E}\left\|\mathbf{x}_{i,u,0} + \sum_{s=0}^{v-1}(\mathbf{e}^x_{i,u,s} - \eta_u\mathbf{O}_{i,u,s}) - \mathbf{X}_u\tfrac{\mathbf{1}}{n}\right\|_F + (1+\sqrt{q^g_{i,u}})L\sqrt{q^x_{i,u}}\mathbb{E}\|\mathbf{x}_{i,u,v}\|_F\Bigg] \\
&\le \frac{1}{\sum_{u=0}^{T-1}\eta_u}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\Bigg[2\sqrt{q^g_{i,u}}\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_* + \sigma_1 + \sigma_2\right) + 2\sigma_1 \\
&\qquad + 2\sqrt{q^g_{i,u}}L\left(\mathbb{E}\|\mathbf{x}_{i,u,0} - \mathbf{X}_u\tfrac{\mathbf{1}}{n}\|_F + \sum_{s=0}^{v-1}\mathbb{E}\|\mathbf{e}^x_{i,u,s}\|_F + \eta_u\sum_{s=0}^{v-1}\mathbb{E}\|\mathbf{O}_{i,u,s}\|_F\right) + (1+\sqrt{q^g_{i,u}})L\sqrt{q^x_{i,u}}\mathbb{E}\|\mathbf{x}_{i,u,v}\|_F\Bigg] \\
&\le \frac{1}{\sum_{u=0}^{T-1}\eta_u}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\sum_{v=0}^{K-1}\Bigg[2\sqrt{q^g_{i,u}}\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_* + \sigma_1 + \sigma_2\right) + 2\sigma_1 \\
&\qquad + 2\sqrt{q^g_{i,u}}L\Bigg(\sum_{\ell=0}^{u-1}\rho^{(u-\ell)/2}\left[\sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q^x_{a,\ell}}\mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K\sqrt{nr}(1+\varepsilon_J)\right] \\
&\qquad\qquad + \sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q^x_{a,u}}\mathbb{E}\|\mathbf{x}_{a,u,j}\|_F + \eta_u K\sqrt{nr}(1+\varepsilon_J)\Bigg) + (1+\sqrt{q^g_{i,u}})L\sqrt{q^x_{i,u}}\mathbb{E}\|\mathbf{x}_{i,u,v}\|_F\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{K}{\sum_{u=0}^{T-1}\eta_u}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=0}^{T-1}\eta_u\Bigg[2\sqrt{q_{i,u}^{g}}\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_u\tfrac{\mathbf{1}}{n})\right\|_* + \sigma_1 + \sigma_2\right) + 2\sigma_1 \\
&\quad + 2\sqrt{q_{i,u}^{g}}L\sum_{\ell=0}^{u}\rho^{(u-\ell)/2}\left[\sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q_{a,\ell}^{x}}\mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K\sqrt{nr}(1+\varepsilon_J)\right] \\
&\quad + (1+\sqrt{q_{i,u}^{g}})L\sqrt{q_{i,u}^{x}}\frac{1}{K}\sum_{v=0}^{K-1}\mathbb{E}\|\mathbf{x}_{i,u,v}\|_F\Bigg] \\
&\le \frac{K}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\Bigg[2\sqrt{q_{i,t}^{g}}\left(\mathbb{E}\left\|\nabla f(\mathbf{X}_t\tfrac{\mathbf{1}}{n})\right\|_* + \sigma_1 + \sigma_2\right) + 2\sigma_1 \\
&\quad + L\left(2\sqrt{q_{i,t}^{g}} + \frac{1+\sqrt{q_{i,t}^{g}}}{K}\right)\sum_{\ell=0}^{t}\rho^{(t-\ell)/2}\left[\sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q_{a,\ell}^{x}}\mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K\sqrt{nr}(1+\varepsilon_J)\right]\Bigg]. \qquad \text{(S.44)}
\end{aligned}
$$

□

Lemma S.10 (Bound on the Momentum Buffer Quantization Error)**.** *Under Assumptions 1–4, Lemmas 2 and 3 of the main text, and Lemmas S.3 and S.2, with the initial consensus condition* $\mathbf{X}_0(\mathbf{I}-\mathbf{J}) = \mathbf{0}$*, for* $0 \le \beta < 1$ *and positive non-increasing stepsizes* $\{\eta_t\}_{t=0}^{T-1}$*, assume*

$$\mathbf{Z}_{i,0,0} = \mathbf{M}_{i,0,0} = \mathbf{M}_{i,0,0}^{Q} = \mathbf{0},$$

and, for $1 \le t < T$*,*

$$\mathbf{Z}_{i,t,0} = \mathbf{Z}_{i,t-1,K}, \qquad \mathbf{M}_{i,t,0} = \mathbf{M}_{i,t-1,K}, \qquad \mathbf{M}_{i,t,0}^{Q} = \mathbf{M}_{i,t-1,K}^{Q}.$$

Then the cumulative momentum-buffer quantization error satisfies

$$
\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{Z}_{i,t,k+1} - \mathbf{M}_{i,t,k+1}^{Q}\|_F \\
&\le \frac{K}{(1-\beta)n\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\left[\sum_{u=t}^{T-1}\beta^{u-t}\sqrt{q_{i,u}^{M}}\prod_{\ell=t}^{u}\left(1+\sqrt{q_{i,\ell}^{M}}\right)^{K}\right]\left(1+\sqrt{q_{i,t}^{g}}\right) \\
&\quad \times\left[\mathbb{E}\left\|\nabla f\left(\mathbf{X}_t\tfrac{\mathbf{1}}{n}\right)\right\|_* + \sigma_1 + \sigma_2 + 2L\sum_{\ell=0}^{t}\rho^{(t-\ell)/2}\left[\sum_{a=1}^{n}\sum_{s=0}^{K}\sqrt{q_{a,\ell}^{x}}\mathbb{E}\|\mathbf{x}_{a,\ell,s}\|_F + \eta_\ell K\sqrt{nr}(1+\varepsilon_J)\right]\right]. \qquad \text{(S.45)}
\end{aligned}
$$

Proof. Using the definitions of $\mathbf{Z}$ and $\mathbf{M}$, we first have

$$\mathbf{Z}_{i,t,k+1} - \mathbf{M}_{i,t,k+1}^{Q} = \beta\left(\mathbf{Z}_{i,t,k} - \mathbf{M}_{i,t,k}^{Q}\right) + \left(\mathbf{M}_{i,t,k+1} - \mathbf{M}_{i,t,k+1}^{Q}\right). \qquad \text{(S.46)}$$

Unrolling Eq. (S.46) over local steps and round boundaries, applying the triangle inequality and the first-moment quantization bound, and then reindexing the source round and local step give

$$
\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{Z}_{i,t,k+1} - \mathbf{M}_{i,t,k+1}^{Q}\|_F \\
&\le \frac{1}{n\sum_{t=0}^{T-1}\eta_t}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\sum_{v=0}^{K-1}\sqrt{q_{i,t}^{M}}\,\mathbb{E}\|\mathbf{M}_{i,t,v+1}\|_F\left[\eta_t\sum_{k=v}^{K-1}\beta^{k-v} + \sum_{s=t+1}^{T-1}\eta_s\sum_{k=0}^{K-1}\beta^{(s-t)K+k-v}\right] \\
&\le \frac{1}{n\sum_{t=0}^{T-1}\eta_t}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\eta_t\sqrt{q_{i,t}^{M}}\sum_{v=0}^{K-1}\mathbb{E}\|\mathbf{M}_{i,t,v+1}\|_F\left[\sum_{k=v}^{K-1}\beta^{k-v} + \sum_{s=t+1}^{T-1}\sum_{k=0}^{K-1}\beta^{(s-t)K+k-v}\right] \\
&\le \frac{1}{(1-\beta)n\sum_{t=0}^{T-1}\eta_t}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\eta_t\sqrt{q_{i,t}^{M}}\sum_{v=0}^{K-1}\mathbb{E}\|\mathbf{M}_{i,t,v+1}\|_F. \qquad \text{(S.47)}
\end{aligned}
$$

The second inequality uses the non-increasing stepsizes $\eta_s \le \eta_t$ for $s \ge t$, while the last one follows from

$$\sum_{k=v}^{K-1} \beta^{k-v} + \sum_{s=t+1}^{T-1} \sum_{k=0}^{K-1} \beta^{(s-t)K+k-v} = \sum_{h=0}^{(T-t)K-1-v} \beta^h \le \frac{1}{1-\beta}.$$

We next directly bound the sum of the momentum-buffer norms, without introducing a separate bound for each $\mathbf{M}_{i,u,v+1}$. Recursively expanding the momentum updates across local steps and round boundaries and summing over $v = 0, \dots, K-1$ yield

$$\begin{aligned}
&\sum_{v=0}^{K-1} \mathbb{E}\|\mathbf{M}_{i,u,v+1}\|_F \\
&\le (1-\beta) \left(\sum_{v=0}^{K-1} \left[\beta \left(1 + \sqrt{q_{i,u}^M} \right) \right]^v \right) \sum_{\tau=0}^{u-1} \sum_{j=0}^{K-1} \left[\beta \left(1 + \sqrt{q_{i,\tau}^M} \right) \right]^{K-j} \\
&\qquad \times \left(\prod_{\ell=\tau+1}^{u-1} \left[\beta \left(1 + \sqrt{q_{i,\ell}^M} \right) \right]^K \right) \mathbb{E}\|\nabla^Q F_i(\mathbf{x}_{i,\tau,j}^Q; \xi_{i,\tau,j})\|_F \\
&\qquad + (1-\beta) \sum_{j=0}^{K-1} \left(\sum_{v=j}^{K-1} \left[\beta \left(1 + \sqrt{q_{i,u}^M} \right) \right]^{v-j} \right) \mathbb{E}\|\nabla^Q F_i(\mathbf{x}_{i,u,j}^Q; \xi_{i,u,j})\|_F \\
&\le \sum_{\tau=0}^{u-1} \beta^{u-\tau} \left(\prod_{\ell=\tau}^{u} \left(1 + \sqrt{q_{i,\ell}^M} \right)^K \right) \sum_{j=0}^{K-1} \mathbb{E}\|\nabla^Q F_i(\mathbf{x}_{i,\tau,j}^Q; \xi_{i,\tau,j})\|_F \\
&\qquad + \left(1 + \sqrt{q_{i,u}^M} \right)^K \sum_{j=0}^{K-1} \mathbb{E}\|\nabla^Q F_i(\mathbf{x}_{i,u,j}^Q; \xi_{i,u,j})\|_F \\
&= \sum_{\tau=0}^{u} \beta^{u-\tau} \left(\prod_{\ell=\tau}^{u} \left(1 + \sqrt{q_{i,\ell}^M} \right)^K \right) \sum_{j=0}^{K-1} \mathbb{E}\|\nabla^Q F_i(\mathbf{x}_{i,\tau,j}^Q; \xi_{i,\tau,j})\|_F. \qquad \text{(S.48)}
\end{aligned}$$

Here we use

$$(1-\beta) \sum_{v=0}^{K-1} \left[\beta(1 + \sqrt{q_{i,u}^M}) \right]^v \le \left(1 + \sqrt{q_{i,u}^M} \right)^{K-1},$$

$$\left[\beta(1 + \sqrt{q_{i,\tau}^M}) \right]^{K-j} \le \beta \left(1 + \sqrt{q_{i,\tau}^M} \right)^K,$$

and, for $u > \tau$,

$$\beta^{1+K(u-\tau-1)} \le \beta^{u-\tau}.$$

For the current-round term, similarly,

$$(1-\beta) \sum_{v=j}^{K-1} \left[\beta(1 + \sqrt{q_{i,u}^M}) \right]^{v-j} \le \left(1 + \sqrt{q_{i,u}^M} \right)^K.$$

Next, for every i, t, j, the quantized stochastic gradient satisfies

$$\begin{aligned}
&\mathbb{E}\|\nabla^Q F_i(\mathbf{x}_{i,t,j}^Q; \xi_{i,t,j})\|_F \\
&\le \left(1 + \sqrt{q_{i,t}^g} \right) \mathbb{E}\|\nabla F_i(\mathbf{x}_{i,t,j}^Q; \xi_{i,t,j})\|_F \\
&\le \left(1 + \sqrt{q_{i,t}^g} \right) \left[\mathbb{E}\left\|\nabla f\left(\mathbf{X}_t \tfrac{\mathbf{1}}{n}\right)\right\|_* + \sigma_1 + \sigma_2 + L\sqrt{q_{i,t}^x}\, \mathbb{E}\|\mathbf{x}_{i,t,j}\|_F + L\mathbb{E}\|\mathbf{x}_{i,t,j} - \mathbf{X}_t \tfrac{\mathbf{1}}{n}\|_F \right] \\
&\le \left(1 + \sqrt{q_{i,t}^g} \right) \Bigg[\mathbb{E}\left\|\nabla f\left(\mathbf{X}_t \tfrac{\mathbf{1}}{n}\right)\right\|_* + \sigma_1 + \sigma_2 \\
&\qquad + 2L \sum_{\ell=0}^{t} \rho^{(t-\ell)/2} \left[\sum_{a=1}^{n} \sum_{s=0}^{K} \sqrt{q_{a,\ell}^x}\, \mathbb{E}\|\mathbf{x}_{a,\ell,s}\|_F + \eta_\ell K \sqrt{nr}(1 + \varepsilon_J) \right] \Bigg]. \qquad \text{(S.49)}
\end{aligned}$$

The last bound uses the parameter-quantization bound, the consensus bound, the intra-round unrolling, and $\rho^0 = 1$ to incorporate the current-round contribution into the same convolution.

Since Eq. (S.49) is uniform over $j = 0, \dots, K-1$, substituting it into Eq. (S.48) gives

$$\begin{aligned}
&\sum_{v=0}^{K-1} \mathbb{E}\|\mathbf{M}_{i,u,v+1}\|_F \\
&\le K \sum_{\tau=0}^{u} \beta^{u-\tau} \left(\prod_{\ell=\tau}^{u} \left(1 + \sqrt{q_{i,\ell}^M}\right)^K \right) \left(1 + \sqrt{q_{i,\tau}^g}\right) \\
&\quad \times \Bigg[\mathbb{E} \left\| \nabla f\left(\mathbf{X}_\tau \tfrac{\mathbf{1}}{n}\right) \right\|_* + \sigma_1 + \sigma_2 \\
&\qquad + 2L \sum_{\ell=0}^{\tau} \rho^{(\tau-\ell)/2} \left[\sum_{a=1}^{n} \sum_{s=0}^{K} \sqrt{q_{a,\ell}^x}\, \mathbb{E}\|\mathbf{x}_{a,\ell,s}\|_F + \eta_\ell K \sqrt{nr}(1+\varepsilon_J) \right] \Bigg]. \qquad \text{(S.50)}
\end{aligned}$$

Finally, substituting Eq. (S.50) into Eq. (S.47) gives

$$\begin{aligned}
&\frac{1}{\sum_{t=0}^{T-1} \eta_t} \sum_{t=0}^{T-1} \frac{\eta_t}{n} \sum_{i=1}^{n} \sum_{k=0}^{K-1} \mathbb{E}\|\mathbf{Z}_{i,t,k+1} - \mathbf{M}_{i,t,k+1}^Q\|_F \\
&\le \frac{K}{(1-\beta) n \sum_{t=0}^{T-1} \eta_t} \sum_{i=1}^{n} \sum_{u=0}^{T-1} \eta_u \sqrt{q_{i,u}^M} \sum_{\tau=0}^{u} \beta^{u-\tau} \left(\prod_{\ell=\tau}^{u} \left(1 + \sqrt{q_{i,\ell}^M}\right)^K \right) \left(1 + \sqrt{q_{i,\tau}^g}\right) \\
&\quad \times \Bigg[\mathbb{E} \left\| \nabla f\left(\mathbf{X}_\tau \tfrac{\mathbf{1}}{n}\right) \right\|_* + \sigma_1 + \sigma_2 \\
&\qquad + 2L \sum_{\ell=0}^{\tau} \rho^{(\tau-\ell)/2} \left[\sum_{a=1}^{n} \sum_{s=0}^{K} \sqrt{q_{a,\ell}^x}\, \mathbb{E}\|\mathbf{x}_{a,\ell,s}\|_F + \eta_\ell K \sqrt{nr}(1+\varepsilon_J) \right] \Bigg] \\
&= \frac{K}{(1-\beta) n \sum_{t=0}^{T-1} \eta_t} \sum_{\tau=0}^{T-1} \sum_{i=1}^{n} \left[\sum_{u=\tau}^{T-1} \eta_u \beta^{u-\tau} \sqrt{q_{i,u}^M} \prod_{\ell=\tau}^{u} \left(1 + \sqrt{q_{i,\ell}^M}\right)^K \right] \left(1 + \sqrt{q_{i,\tau}^g}\right) \\
&\quad \times \Bigg[\mathbb{E} \left\| \nabla f\left(\mathbf{X}_\tau \tfrac{\mathbf{1}}{n}\right) \right\|_* + \sigma_1 + \sigma_2 \\
&\qquad + 2L \sum_{\ell=0}^{\tau} \rho^{(\tau-\ell)/2} \left[\sum_{a=1}^{n} \sum_{s=0}^{K} \sqrt{q_{a,\ell}^x}\, \mathbb{E}\|\mathbf{x}_{a,\ell,s}\|_F + \eta_\ell K \sqrt{nr}(1+\varepsilon_J) \right] \Bigg] \\
&\le \frac{K}{(1-\beta) n \sum_{t=0}^{T-1} \eta_t} \sum_{t=0}^{T-1} \eta_t \sum_{i=1}^{n} \left[\sum_{u=t}^{T-1} \beta^{u-t} \sqrt{q_{i,u}^M} \prod_{\ell=t}^{u} \left(1 + \sqrt{q_{i,\ell}^M}\right)^K \right] \left(1 + \sqrt{q_{i,t}^g}\right) \\
&\quad \times \left[\mathbb{E} \left\| \nabla f\left(\mathbf{X}_t \tfrac{\mathbf{1}}{n}\right) \right\|_* + \sigma_1 + \sigma_2 + 2L \sum_{\ell=0}^{t} \rho^{(t-\ell)/2} \left[\sum_{a=1}^{n} \sum_{s=0}^{K} \sqrt{q_{a,\ell}^x}\, \mathbb{E}\|\mathbf{x}_{a,\ell,s}\|_F + \eta_\ell K \sqrt{nr}(1+\varepsilon_J) \right] \right]. \qquad \text{(S.51)}
\end{aligned}$$

where the last inequality uses $\eta_u \le \eta_t$ for $u \ge t$. This completes the proof. □

A.2 PROOF OF THEOREM 1

Proof. Let $S_1 \triangleq \sum_{t=0}^{T-1} \eta_t$ and $S_2 \triangleq \sum_{t=0}^{T-1} \eta_t^2$, with $r = \min\{m, d_{\mathrm{m}}\}$. Under Assumptions 1–4, Lemmas 2 and 3 of the main text, and Lemmas S.1, S.2, S.3, S.4, S.5, S.6, S.7, S.8, S.9, and S.10, provided that the coefficient multiplying $\mathbb{E}\left\|\nabla f(\mathbf{X}_t \frac{\mathbf{1}}{n})\right\|_*$ in the intermediate bound is positive for every $t \in \{0, \dots, T-1\}$ (equivalently $\Delta > 0$, see Eq. (S.57)), expanding the expected objective function value $\mathbb{E}\left[f\left(\mathbf{X}_{t+1} \frac{\mathbf{1}}{n}\right)\right]$ at round $t+1$ yields the following single continuous step-by-step

descent chain:

$$
\begin{aligned}
&\mathbb{E}\left[f\left(\mathbf{X}_{t+1}\frac{\mathbf{1}}{n}\right)\right] \\
&\leq \mathbb{E}\left[f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right] + \mathbb{E}\left\langle \nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right), \sum_{k=0}^{K}\mathbf{e}_{t,k}^{x}\frac{\mathbf{1}}{n} - \eta_t\sum_{k=0}^{K-1}\mathbf{O}_{t,k}\frac{\mathbf{1}}{n}\right\rangle_F \\
&\quad + \frac{L}{2}\mathbb{E}\|\sum_{k=0}^{K}\mathbf{e}_{t,k}^{x}\frac{\mathbf{1}}{n} - \eta_t\sum_{k=0}^{K-1}\mathbf{O}_{t,k}\frac{\mathbf{1}}{n}\|_F^2 \\
&\leq \mathbb{E}\left[f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right] + \mathbb{E}\left\langle \nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right), \sum_{k=0}^{K}\mathbf{e}_{t,k}^{x}\frac{\mathbf{1}}{n}\right\rangle_F \\
&\quad - \eta_t\mathbb{E}\left\langle \nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right), \sum_{k=0}^{K-1}\mathbf{O}_{t,k}\frac{\mathbf{1}}{n}\right\rangle_F + L\mathbb{E}\|\sum_{k=0}^{K}\mathbf{e}_{t,k}^{x}\frac{\mathbf{1}}{n}\|_F^2 + L\eta_t^2\mathbb{E}\|\sum_{k=0}^{K-1}\mathbf{O}_{t,k}\frac{\mathbf{1}}{n}\|_F^2 \\
&\leq \mathbb{E}\left[f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right] - \frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\left\langle \nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right), \mathbf{O}_{i,t,k}\right\rangle_F \\
&\quad + \frac{L(K+1)}{n}\sum_{i=1}^{n}\sum_{k=0}^{K}\mathbb{E}\|\mathbf{e}_{i,t,k}^{x}\|_F^2 + \frac{LK\eta_t^2}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{O}_{i,t,k}\|_F^2 \\
&\leq \mathbb{E}\left[f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right] - \frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\langle \nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right), \mathbf{P}_{i,t,k}\right\rangle_F + \left\langle \nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right), \mathbf{O}_{i,t,k} - \mathbf{P}_{i,t,k}\right\rangle_F\right] \\
&\quad + \frac{L(K+1)}{n}\sum_{i=1}^{n}\sum_{k=0}^{K}q_{i,t}^{x}\mathbb{E}\|\mathbf{x}_{i,t,k}\|_F^2 + LrK^2\eta_t^2(1+\varepsilon_J)^2 \\
&\leq \mathbb{E}\left[f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right] - \frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\Bigg[\left\langle \mathbf{M}_{i,t,k+1}^{Q}, \mathbf{P}_{i,t,k}\right\rangle_F \\
&\qquad - \left\langle \mathbf{M}_{i,t,k+1}^{Q} - \nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right), \mathbf{P}_{i,t,k}\right\rangle_F + \left\langle \nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right), \mathbf{O}_{i,t,k} - \mathbf{P}_{i,t,k}\right\rangle_F\Bigg] \\
&\quad + \frac{L(K+1)}{n}\sum_{i=1}^{n}\sum_{k=0}^{K}q_{i,t}^{x}\mathbb{E}\|\mathbf{x}_{i,t,k}\|_F^2 + LrK^2\eta_t^2(1+\varepsilon_J)^2 \\
&\leq \mathbb{E}\left[f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right] - \frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbf{M}_{i,t,k+1}^{Q}\right\|_* - \left\|\mathbf{M}_{i,t,k+1}^{Q} - \nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right\|_*\|\mathbf{P}_{i,t,k}\|_{\mathrm{op}} - \left\|\nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right\|_*\|\mathbf{O}_{i,t,k} - \mathbf{P}_{i,t,k}\|_{\mathrm{op}}\right] \\
&\quad + \frac{L(K+1)}{n}\sum_{i=1}^{n}\sum_{k=0}^{K}q_{i,t}^{x}\mathbb{E}\|\mathbf{x}_{i,t,k}\|_F^2 + LrK^2\eta_t^2(1+\varepsilon_J)^2 \\
&\leq \mathbb{E}\left[f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right] - \frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\left[(1-\varepsilon_J)\left\|\nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right\|_* - 2\left\|\mathbf{M}_{i,t,k+1}^{Q} - \nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right\|_*\right] \\
&\quad + \frac{L(K+1)}{n}\sum_{i=1}^{n}\sum_{k=0}^{K}q_{i,t}^{x}\mathbb{E}\|\mathbf{x}_{i,t,k}\|_F^2 + LrK^2\eta_t^2(1+\varepsilon_J)^2 \\
&\leq \mathbb{E}\left[f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right] - K\eta_t(1-\varepsilon_J)\mathbb{E}\left\|\nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\right\|_* \\
&\quad + \frac{2\sqrt{r}\,\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{M}_{i,t,k+1}^{Q} - \nabla f\left(\mathbf{X}_{t}\frac{\mathbf{1}}{n}\right)\|_F \\
&\quad + \frac{L(K+1)}{n}\sum_{i=1}^{n}\sum_{k=0}^{K}q_{i,t}^{x}\mathbb{E}\|\mathbf{x}_{i,t,k}\|_F^2 + LrK^2\eta_t^2(1+\varepsilon_J)^2. \qquad \text{(S.52)}
\end{aligned}
$$

The two displayed inequalities that bound the bracket from below (the bracket enters with the negative prefactor $-\eta_t/n$) use polar duality $\left\langle \mathbf{M}^Q_{i,t,k+1}, \mathbf{P}_{i,t,k} \right\rangle_F = \left\| \mathbf{M}^Q_{i,t,k+1} \right\|_*$, Hölder's inequality, $\|\mathbf{P}_{i,t,k}\|_{\mathrm{op}} \le 1$, the reverse triangle inequality, and the approximation bound $\|\mathbf{O}_{i,t,k} - \mathbf{P}_{i,t,k}\|_{\mathrm{op}} \le \varepsilon_J$ from Lemma 3. The linear forward-quantization terms in this chain vanish by the conditional unbiasedness of the quantizer (Lemma 2). To bound the tracking-error term in Eq. (S.52), we first decompose the tracking error into six terms, substitute the corresponding cumulative bounds from Lemmas S.6, S.7, S.8, S.9, and S.10, together with Lemma S.5. For notational simplicity, we define the cumulative momentum-quantization propagation factor from round t for client i as

$$\mathcal{U}_{i,t} \triangleq \sum_{u=t}^{T-1} \beta^{u-t} \sqrt{q^M_{i,u}} \prod_{\ell=t}^{u} \left(1 + \sqrt{q^M_{i,\ell}}\right)^K. \tag{S.53}$$

$$\begin{aligned}
&\frac{1}{S_1}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{M}^Q_{i,t,k+1} - \nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)\|_F \\
&\le \frac{1}{S_1}\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Bigg[\mathbb{E}\|\mathbf{M}^Q_{i,t,k+1} - \mathbf{Z}_{i,t,k+1}\|_F + \mathbb{E}\|\mathbf{Z}_{i,t,k+1} - \mathbf{Y}_{i,t,k+1}\|_F + \mathbb{E}\|\mathbf{Y}_{i,t,k+1} - \mathbf{S}_{i,t,k+1}\|_F \\
&\qquad + \mathbb{E}\|\mathbf{S}_{i,t,k+1} - \mathbf{C}_{i,t,k+1}\|_F + \mathbb{E}\|\mathbf{C}_{i,t,k+1} - \nabla f_i(\mathbf{x}_{i,t,k})\|_F \\
&\qquad + \mathbb{E}\|\nabla f_i(\mathbf{x}_{i,t,k}) - \nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)\|_F\Bigg] \\
&\le K\sigma_1\sqrt{\frac{1-\beta}{1+\beta}} + 2K\sigma_1 + K\sigma_2 + 2\beta K\sigma_2 \\
&\quad + LK\sqrt{r}(1+\varepsilon_J)\left[\frac{K}{1-\sqrt{\rho}} + \frac{2\beta K}{1-\sqrt{\rho}} + \frac{\beta}{1-\beta}\right]\frac{S_2}{S_1} \\
&\quad + \frac{L}{S_1}\left[\frac{K}{\sqrt{n}(1-\sqrt{\rho})} + \frac{2\beta K}{(1-\beta)\sqrt{n}(1-\sqrt{\rho})} + \frac{\beta}{n(1-\beta)}\right]\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q^x_{i,t}}\,\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F \\
&\quad + \frac{1}{nS_1}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\Bigg\{\left[\frac{\beta}{1-\beta} + 3K\sqrt{q^g_{i,t}} + \frac{K}{1-\beta}\mathcal{U}_{i,t}\left(1+\sqrt{q^g_{i,t}}\right)\right]\mathbb{E}\left\|\nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)\right\|_* \\
&\qquad + 3K\sqrt{q^g_{i,t}}(\sigma_1+\sigma_2) \\
&\qquad + L\left[1 + (3K+1)\sqrt{q^g_{i,t}}\right]\sum_{\ell=0}^{t}\rho^{(t-\ell)/2}\left[\sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q^x_{a,\ell}}\,\mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K\sqrt{nr}(1+\varepsilon_J)\right] \\
&\qquad + \frac{K}{1-\beta}\mathcal{U}_{i,t}\left(1+\sqrt{q^g_{i,t}}\right)\left[\sigma_1+\sigma_2 + 2L\sum_{\ell=0}^{t}\rho^{(t-\ell)/2}\left[\sum_{a=1}^{n}\sum_{j=0}^{K}\sqrt{q^x_{a,\ell}}\,\mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K\sqrt{nr}(1+\varepsilon_J)\right]\right]\Bigg\} \\
&\le 3K(\sigma_1+\sigma_2) + \frac{4LK^2\sqrt{r}(1+\varepsilon_J)}{(1-\beta)(1-\sqrt{\rho})}\frac{S_2}{S_1} \\
&\quad + \frac{3LK}{\sqrt{n}(1-\beta)(1-\sqrt{\rho})S_1}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q^x_{i,t}}\,\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F \\
&\quad + \frac{1}{nS_1}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\Bigg\{\left[\frac{\beta}{1-\beta} + 3K\sqrt{q^g_{i,t}} + \frac{K}{1-\beta}\mathcal{U}_{i,t}\left(1+\sqrt{q^g_{i,t}}\right)\right]\mathbb{E}\left\|\nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)\right\|_* \\
&\qquad + \left[3K\sqrt{q^g_{i,t}} + \frac{K}{1-\beta}\mathcal{U}_{i,t}\left(1+\sqrt{q^g_{i,t}}\right)\right](\sigma_1+\sigma_2) \\
&\qquad + 2LK\left(1+\sqrt{q^g_{i,t}}\right)\left[2 + \frac{\mathcal{U}_{i,t}}{1-\beta}\right]
\end{aligned}$$

$$\times \sum_{\ell=0}^{t} \rho^{(t-\ell)/2} \left[\sum_{a=1}^{n} \sum_{j=0}^{K} \sqrt{q_{a,\ell}^{x}}\, \mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K \sqrt{nr}(1+\varepsilon_J) \right] \Bigg\}. \tag{S.54}$$

Let

$$\bar{\eta} \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \eta_t. \tag{S.55}$$

Substituting Eq. (S.54) into the sum of Eq. (S.52) over $t = 0, \ldots, T-1$ and dividing by $KT\bar{\eta}(1-\varepsilon_J)$ gives the following intermediate bound:

$$\begin{aligned}
&\frac{1}{T\bar{\eta}} \sum_{t=0}^{T-1} \eta_t \mathbb{E} \left\| \nabla f\left(\mathbf{X}_t \frac{\mathbf{1}}{n} \right) \right\|_* \\
&\leq \frac{\mathbb{E}\left[f\left(\mathbf{X}_0 \frac{\mathbf{1}}{n} \right) \right] - f^\star}{KT\bar{\eta}(1-\varepsilon_J)} + \frac{6\sqrt{r}}{1-\varepsilon_J}(\sigma_1 + \sigma_2) \\
&\quad + \frac{LrK}{1-\varepsilon_J} \left[(1+\varepsilon_J)^2 + \frac{8(1+\varepsilon_J)}{(1-\beta)(1-\sqrt{\rho})} \right] \frac{\sum_{t=0}^{T-1} \eta_t^2}{T\bar{\eta}} \\
&\quad + \frac{6L\sqrt{r}}{\sqrt{n}(1-\varepsilon_J)(1-\beta)(1-\sqrt{\rho})T\bar{\eta}} \sum_{t=0}^{T-1} \eta_t \sum_{i=1}^{n} \sum_{j=0}^{K} \sqrt{q_{i,t}^{x}}\, \mathbb{E}\|\mathbf{x}_{i,t,j}\|_F \\
&\quad + \frac{L(K+1)}{nK(1-\varepsilon_J)T\bar{\eta}} \sum_{t=0}^{T-1} \sum_{i=1}^{n} \sum_{k=0}^{K} q_{i,t}^{x} \mathbb{E}\|\mathbf{x}_{i,t,k}\|_F^2 \\
&\quad + \frac{2\sqrt{r}}{K(1-\varepsilon_J)nT\bar{\eta}} \sum_{t=0}^{T-1} \eta_t \sum_{i=1}^{n} \Bigg\{ \left[\frac{\beta}{1-\beta} + 3K\sqrt{q_{i,t}^{g}} + \frac{K}{1-\beta} \mathcal{U}_{i,t} \left(1 + \sqrt{q_{i,t}^{g}} \right) \right] \mathbb{E} \left\| \nabla f\left(\mathbf{X}_t \frac{\mathbf{1}}{n} \right) \right\|_* \\
&\qquad + \left[3K\sqrt{q_{i,t}^{g}} + \frac{K}{1-\beta} \mathcal{U}_{i,t} \left(1 + \sqrt{q_{i,t}^{g}} \right) \right] (\sigma_1 + \sigma_2) \\
&\qquad + 2LK \left(1 + \sqrt{q_{i,t}^{g}} \right) \left[2 + \frac{\mathcal{U}_{i,t}}{1-\beta} \right] \sum_{\ell=0}^{t} \rho^{(t-\ell)/2} \left[\sum_{a=1}^{n} \sum_{j=0}^{K} \sqrt{q_{a,\ell}^{x}}\, \mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K \sqrt{nr}(1+\varepsilon_J) \right] \Bigg\}.
\end{aligned} \tag{S.56}$$

Define

$$\Delta \triangleq \min_{0 \leq t < T} \left\{ 1 - \frac{2\sqrt{r}}{K(1-\varepsilon_J)n} \sum_{i=1}^{n} \left[\frac{\beta}{1-\beta} + 3K\sqrt{q_{i,t}^{g}} + \frac{K}{1-\beta} \mathcal{U}_{i,t} \left(1 + \sqrt{q_{i,t}^{g}} \right) \right] \right\}, \tag{S.57}$$

and assume $\Delta > 0$. Moving the gradient-norm term in Eq. (S.56) to the left-hand side gives

$$\begin{aligned}
&\frac{\Delta}{T\bar{\eta}} \sum_{t=0}^{T-1} \eta_t \mathbb{E} \left\| \nabla f\left(\mathbf{X}_t \frac{\mathbf{1}}{n} \right) \right\|_* \\
&\leq \frac{\mathbb{E}\left[f\left(\mathbf{X}_0 \frac{\mathbf{1}}{n} \right) \right] - f^\star}{KT\bar{\eta}(1-\varepsilon_J)} + \frac{6\sqrt{r}}{1-\varepsilon_J}(\sigma_1 + \sigma_2) \\
&\quad + \frac{LrK}{1-\varepsilon_J} \left[(1+\varepsilon_J)^2 + \frac{8(1+\varepsilon_J)}{(1-\beta)(1-\sqrt{\rho})} \right] \frac{\sum_{t=0}^{T-1} \eta_t^2}{T\bar{\eta}} \\
&\quad + \frac{6L\sqrt{r}}{\sqrt{n}(1-\varepsilon_J)(1-\beta)(1-\sqrt{\rho})T\bar{\eta}} \sum_{t=0}^{T-1} \eta_t \sum_{i=1}^{n} \sum_{j=0}^{K} \sqrt{q_{i,t}^{x}}\, \mathbb{E}\|\mathbf{x}_{i,t,j}\|_F \\
&\quad + \frac{L(K+1)}{nK(1-\varepsilon_J)T\bar{\eta}} \sum_{t=0}^{T-1} \sum_{i=1}^{n} \sum_{k=0}^{K} q_{i,t}^{x} \mathbb{E}\|\mathbf{x}_{i,t,k}\|_F^2 \\
&\quad + \frac{2\sqrt{r}}{K(1-\varepsilon_J)nT\bar{\eta}} \sum_{t=0}^{T-1} \eta_t \sum_{i=1}^{n} \Bigg\{ \left[3K\sqrt{q_{i,t}^{g}} + \frac{K}{1-\beta} \mathcal{U}_{i,t} \left(1 + \sqrt{q_{i,t}^{g}} \right) \right] (\sigma_1 + \sigma_2) \\
&\qquad + 2LK \left(1 + \sqrt{q_{i,t}^{g}} \right) \left[2 + \frac{\mathcal{U}_{i,t}}{1-\beta} \right] \sum_{\ell=0}^{t} \rho^{(t-\ell)/2} \left[\sum_{a=1}^{n} \sum_{j=0}^{K} \sqrt{q_{a,\ell}^{x}}\, \mathbb{E}\|\mathbf{x}_{a,\ell,j}\|_F + \eta_\ell K \sqrt{nr}(1+\varepsilon_J) \right] \Bigg\}.
\end{aligned} \tag{S.58}$$

Dividing both sides of Eq. (S.58) by $\Delta > 0$, and using the definition of Δ together with the non-increasing stepsizes, we obtain

$$\begin{aligned}
&\frac{1}{T\bar{\eta}}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)\right\|_* \\
&\leq \frac{\mathbb{E}\left[f\left(\mathbf{X}_0\frac{\mathbf{1}}{n}\right)\right]-f^\star}{\Delta KT\bar{\eta}(1-\varepsilon_J)}+\left[\frac{6\sqrt{r}}{(1-\varepsilon_J)\Delta}+\frac{1-\Delta}{\Delta}\right](\sigma_1+\sigma_2) \\
&\quad+\left\{\frac{LrK}{(1-\varepsilon_J)\Delta}\left[(1+\varepsilon_J)^2+\frac{8(1+\varepsilon_J)}{(1-\beta)(1-\sqrt{\rho})}\right]+\frac{10LKr\sqrt{n}(1+\varepsilon_J)}{(1-\varepsilon_J)(1-\sqrt{\rho})\Delta}\right\}\frac{\sum_{t=0}^{T-1}\eta_t^2}{T\bar{\eta}} \\
&\quad+\left[\frac{6L\sqrt{r}}{\sqrt{n}(1-\varepsilon_J)(1-\beta)(1-\sqrt{\rho})\Delta}+\frac{10L\sqrt{r}}{(1-\varepsilon_J)(1-\sqrt{\rho})\Delta}\right]\frac{1}{T\bar{\eta}}\sum_{t=0}^{T-1}\eta_t\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^x}\,\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F \\
&\quad+\frac{L(K+1)}{\Delta nK(1-\varepsilon_J)T\bar{\eta}}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\sum_{k=0}^{K}q_{i,t}^x\mathbb{E}\|\mathbf{x}_{i,t,k}\|_F^2. \qquad \text{(S.59)}
\end{aligned}$$

Choose

$$\eta_t = D_t\sqrt{\frac{1-\beta}{KT}}, \qquad D_m = \min_{0\leq t<T} D_t, \qquad D_M = \max_{0\leq t<T} D_t. \qquad \text{(S.60)}$$

with $\{D_t\}_{t=0}^{T-1}$ non-increasing, and let

$$\beta = 1-\min\left\{\sqrt{\frac{nL}{rKT}},1\right\}. \qquad \text{(S.61)}$$

Substituting these bounds into Eq. (S.59) gives

$$\begin{aligned}
&\frac{1}{T\bar{\eta}}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\left\|\nabla f\left(\mathbf{X}_t\frac{\mathbf{1}}{n}\right)\right\|_* \\
&\leq \frac{r^{1/4}\left(\mathbb{E}\left[f\left(\mathbf{X}_0\frac{\mathbf{1}}{n}\right)\right]-f^\star\right)}{\Delta D_m(1-\varepsilon_J)n^{1/4}L^{1/4}K^{1/4}T^{1/4}} \\
&\quad+\left[\frac{6\sqrt{r}}{(1-\varepsilon_J)\Delta}+\frac{1-\Delta}{\Delta}\right](\sigma_1+\sigma_2) \\
&\quad+\left\{\frac{LrK}{(1-\varepsilon_J)\Delta}\left[(1+\varepsilon_J)^2+\frac{8(1+\varepsilon_J)}{1-\sqrt{\rho}}\sqrt{\frac{rKT}{nL}}\right]\right. \\
&\qquad\left.+\frac{10LKr\sqrt{n}(1+\varepsilon_J)}{(1-\varepsilon_J)(1-\sqrt{\rho})\Delta}\right\}\frac{D_M n^{1/4}L^{1/4}}{r^{1/4}K^{3/4}T^{3/4}} \\
&\quad+\left[\frac{6L\sqrt{r}}{\sqrt{n}(1-\varepsilon_J)(1-\sqrt{\rho})\Delta}\sqrt{\frac{rKT}{nL}}+\frac{10L\sqrt{r}}{(1-\varepsilon_J)(1-\sqrt{\rho})\Delta}\right] \\
&\qquad\times\frac{D_M}{D_mT}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\sum_{j=0}^{K}\sqrt{q_{i,t}^x}\,\mathbb{E}\|\mathbf{x}_{i,t,j}\|_F \\
&\quad+\frac{L^{3/4}r^{1/4}(K+1)K^{-1/4}}{\Delta D_m(1-\varepsilon_J)n^{5/4}T^{1/4}}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\sum_{k=0}^{K}q_{i,t}^x\mathbb{E}\|\mathbf{x}_{i,t,k}\|_F^2. \qquad \text{(S.62)}
\end{aligned}$$

□

B EXPERIMENTAL DETAILS

Common setup and baselines. For image classification, we use ResNet-18 on TinyImageNet-200 Le et al. (2015). For language-model pretraining, we use nanoGPT Karpathy (2023) on FineWeb-Edu Penedo et al. (2024). For language-model fine-tuning, we use GPT-2 XL with LoRA on the GLUE benchmark Wang et al. (2018).

To evaluate *RADeMuon*, we compare it with the following baselines:

- **DPSGD Lian et al. (2017):** Each client performs local SGD updates in full precision and exchanges model parameters through the gossip matrix.
- **LM-DFL Chen et al. (2024):** It applies Lloyd–Max quantization to communicated model parameters while retaining full-precision local SGD updates.
- **LPT Li et al. (2024):** It is based on DPSGD and quantizes the numerical quantities involved in local optimization, including gradients, optimizer states, and model weights, together with the model parameters exchanged during communication.
- **Adriana Qiao et al. (2026):** Each client adaptively selects its training quantization scale according to its available computational resources, performs low-precision local training, and then adaptively selects a communication quantization scale based on its communication budget before transmitting quantized model parameters to neighboring clients.
- **De-EF21-Muon Gruntkowska et al. (2026):** EF21-Muon performs Muon updates from an aggregated gradient estimator using a linear minimization oracle, with error-feedback compression of model updates and gradient or momentum increments. De-EF21-Muon replaces its server aggregation with topology-weighted peer-to-peer mixing and quantizes the communicated model parameters.

We consider 5 clients connected by a randomly generated doubly stochastic communication matrix, with weights constructed using the Metropolis–Hastings procedure Awan et al. (2006). Unless otherwise specified, full precision refers to 32-bit floating-point precision.

For *RADeMuon*, every quantized tensor is divided into blocks of at most 256 entries. The supplied precision configuration is 32 bits for rounds $0 \leq t < T/6$, 16 bits for $T/6 \leq t < T/3$, and 8 bits for the remaining rounds. The same schedule is applied to parameter, gradient, and momentum representations in the current experiments. Following Adriana (Qiao et al., 2026), both resource metrics are reported in MB. Comp. is the average local computation cost per client, whereas Comm. is the total communication volume of all five clients over all communication rounds and accounts for the topology used in each round.

All experiments are conducted with PyTorch 2.10.0, Python 3.12, and CUDA 12.8 on a server running Ubuntu 22.04, equipped with four NVIDIA RTX 5090 GPUs and four NVIDIA RTX 3090 GPUs.

B.1 IMAGE CLASSIFICATION

We train ResNet-18 on TinyImageNet-200 (Le et al., 2015), which contains 200 classes of 64×64 images. We partition labels with Dirichlet concentration $\alpha \in \{0.1, 0.5\}$. Each run lasts 300 communication rounds; every client performs 50 local steps per round with batch size 128. SGD-based methods use learning rate 10^{-3}. The Muon-based methods use Muon and auxiliary AdamW learning rates of 0.015, gradient clipping at 10, momentum $\beta = 0.95$, and 5 NS steps. We report the Acc@1/Acc@5. The results appear in Table 2.

B.2 LANGUAGE-MODEL PRETRAINING

We pretrain a nanoGPT model (Karpathy, 2023) on FineWeb-Edu (Penedo et al., 2024). The model has 12 layers, 12 attention heads, hidden size 768, feed-forward size 3072, a vocabulary of 50,304 tokens, tied embeddings, RoPE position embeddings (Su et al., 2024), approximately 124M parameters, and context length 1,024. The training budget is 2.5B tokens, approximately 20 tokens per parameter following compute-optimal scaling (Hoffmann et al., 2022). Each client performs 30 local steps per round with 8 sequences per batch. We use a WSD schedule with 500 warm-up steps and

Table 4: Full GLUE fine-tuning results. All reported values use two decimal places. Comp. is the average local computation cost per client; Comm. is the total communication volume of all five clients over all rounds.

Method	Bit	GLUE tasks								Efficiency	
		CoLA	RTE	SST-2	QQP	MRPC	QNLI	MNLI	Avg.	Comp.(MB)	Comm.(MB)
DPSGD	32	59.33	76.53	94.95	81.31	78.73	89.72	80.59	80.17	75.00	862.64
LM-DFL	32	59.14	77.62	94.95	81.31	79.41	89.74	80.59	80.39	75.00	1725.28
	16	58.38	75.88	94.95	81.31	79.66	89.72	80.60	80.07	75.00	862.64
	8	55.80	71.91	95.23	81.30	77.75	89.73	72.95	77.81	75.00	431.32
LPT	32	58.83	75.88	94.95	81.31	79.26	89.73	80.61	80.08	75.00	862.64
	16	56.37	78.34	94.95	81.26	79.85	89.69	80.65	80.16	37.50	431.32
	8	54.13	75.88	94.04	80.11	82.35	88.17	79.48	79.17	18.75	215.66
Adriana	32	57.30	78.34	95.18	81.28	80.83	89.75	80.76	80.49	75.00	301.24
	16	57.26	78.34	95.16	81.25	80.74	89.71	80.75	80.46	37.50	301.29
	8	52.67	74.15	94.40	80.66	79.26	89.31	79.42	78.55	18.75	301.31
De-EF21-Muon	32	56.62	76.17	95.09	81.85	79.17	89.63	80.61	79.88	75.00	862.64
	16	56.56	75.81	95.07	81.85	79.17	89.63	80.61	79.81	75.00	461.78
	8	55.56	75.09	95.05	82.07	78.19	89.51	80.62	79.44	75.00	246.12
RADeMuon	–	56.56	75.45	95.07	81.85	78.92	89.63	80.61	79.73	31.13	356.61

decay to 0.15 times the peak learning rate. SGD-based methods use learning rate 6×10^{-4}. Muon-based methods use Muon learning rate 0.02, auxiliary AdamW learning rate 6×10^{-4}, gradient clipping at 10, momentum $\beta = 0.95$, 5 NS steps, and Muon-branch weight decay 0.01.

We select the checkpoint with the lowest validation cross-entropy and run zero-shot evaluation with the LM Evaluation Harness (Gao et al., 2021). The 12 tasks are HellaSwag (Zellers et al., 2019), ARC-Easy and ARC-Challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SciQ (Welbl et al., 2017), LAMBADA (Paperno et al., 2016), MMLU (Hendrycks et al., 2020), WinoGrande (Sakaguchi et al., 2021), CommonsenseQA (Talmor et al., 2019), LogiQA (Liu et al., 2020), and COPA (Roemmele et al., 2011).

B.3 Language-Model Fine-Tuning

We fine-tune GPT-2 XL (1.5B parameters) with LoRA on SST-2 (Socher et al., 2013), CoLA (Warstadt et al., 2019), MRPC (Dolan & Brockett, 2005), RTE (Dagan et al., 2005), QNLI (Rajpurkar et al., 2016), MNLI (Williams et al., 2018), and QQP from GLUE (Wang et al., 2018). LoRA uses rank 16, scaling 32, and dropout 0.05 on attention and MLP projections and the classification head, yielding approximately 19.66M trainable parameters. MNLI uses the validation-matched split. We partition data with Dirichlet concentration $\alpha = 0.5$.

Training lasts 100 communication rounds. Each client performs up to 20 local steps with batch size 16, capped at one local epoch. DPSGD, LM-DFL, and LPT use SGD with learning rate 10^{-4}. Muon-based methods use Muon learning rate 3×10^{-4}, auxiliary AdamW learning rate 3×10^{-3}, gradient clipping at 10, momentum $\beta = 0.95$, 5 NS steps, and Muon weight decay 0.01.